مباحث لغوية (١٤)

المدونات اللغوية العربية

بناؤها وطرائق الإفادة منها

تأليف
أ.د.محمود إسماعيل صالح
أ.عبدالله يحيى الفيفي
د.عبدالمحسن عبيد الثبيتي
د.عقيل حامد الشمري
د.سلطان ناصر المجيول

تحرير
د. صالح بن فهد العصيمي

الطبعة الأولى
الرياض
١٤٣٦ هـ - ٢٠١٥م

فهرسة مكتبة الملك فهد الوطنية أثناء النشر

العصيمي، صالح الفهد

المدونات اللغوية العربية بناؤها وطرائق الإفادة منها. /صالح فهد العصيمي .- الرياض، ١٤٣٦ هـ

٣٠٠ ص؛ ١٧ × ٢٤ سم

ردمك: ٤ - ٨ - ٩٠٦٦٤ - ٦٠٣ - ٩٧٨

١- اللغات - تاريخ ٢- اللغة العربية - بحوث أ. العنوان

ديوي ٤١٠,٧٢ ١٤٣٦/٥٣٣٨

رقم الإيداع: ١٤٣٦/٥٣٣٨

ردمك: ٤ - ٨ - ٩٠٦٦٤ - ٦٠٣ - ٩٧٨

الطبعة الأولى

١٤٣٦هـ / ٢٠١٥م

سلسلة من الإصدارات التي تعالج قضايا لغوية متنوعة

مدير المشروع:

أ. خالد بن أحمد الرفاعي

إشراف:

د. عبدالله بن صالح الوشمي



كلمة المركز

يجتهد مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية في العمل في مجالات متعددة تحقق تعميق الوعي اللغوي على المستويات المختلفة (الاجتماعية والعلمية/ الأهلية والرسمية) ؛ وذلك للسمو باللغة العربية، وترسيخ منافستها للغات الحضارية في العالم، وتعميق قيادتها الدينية والتاريخية لشعوب شتى في أنحاء المعمورة.

وامتدادا لذلك. ينشط المركز في مجال النشر، مستقطباً الأعمال العلمية الجادة وفق لائحة معتمدة منظّمة لذلك، كما ينشط في مجال التأليف من خلال استكتاب مجموعة كبيرة من الباحثين ؛ لتأليف عدد متنوع من الإصدارات النوعية المقروءة التي تعالج عنواناتٍ يقتنصها المركز، ويلفت الانتباه إليها، ويعلن من خلالها الفرص الممكنة لخدمة اللغة العربية في المجالات المختلفة، ملبّياً بذلك الحاجات التي يلمس المركز تطلّع المكتبة اللغوية العربية إليها، ولافتاً الأنظار إلى أهمية التعمق فيها بحثياً، واستكشاف ما يمكن عمله تنفيذياً في هذه المجالات. ويسعد المركز بأن استقطب في المرحلة الأولى من هذا المشروع ما يربو على مئتي باحث، موسّعاً دائرة المشاركة محلياً وخليجيا وعربيا وإسلاميا وعالميا، ومنوّعاً مسارات البحث الرئيسية والفرعية، ومنفتحا على كل ما من شأنه خدمة اللغة العربية بجميع الوسائل والأطر.

ويمثّل هذا الكتاب واحدا من الكتب التي صدرت ضمن سلسلة (مباحث لغوية) يحتوي عددا من الأبحاث لأساتذة مرموقين؛ استجابوا لما رآه المركز من الحاجة إلى التأليف تحت هذا العنوان، وبادروا إلى ذلك مشكورين.

وتودّ الأمانة العامة أن تشيد بجهد السادة المؤلفين، وجهد محرر الكتاب، ومدير هذا المشروع العلمي على ما تفضلوا به من التزام علمي لا يستغرب من مثلهم، وقد ترك المركز للمحرر مساحة واسعة من الحرية في اختيار الباحثين

ووضع الخطة العلمية – بالتشاور مع المركز –؛ سعياً إلى تحقيق أقصى ما يمكن تحقيقه من الإفادة العلمية، مع الأخذ بالاعتبار أن الآراء الواردة في البحوث لا تمثل رأي المركز بالضرورة، ولكنها من جملة الآراء العلمية التي يسعد المركز بإتاحتها للمجتمع العلمي وللمعنيين بالشأن اللغوي لتداول الرأي ، وتعميق النظر، ونلفت انتباه القارئ الكريم إلى أن ترتيب أسماء المؤلفين على الغلاف موافق لترتيب أبحاثهم في الكتاب، وهي خاضعة للرؤية المنهجية التي تفضّل المحرر – مشكورا– باقتراح خطتها.

والشكر والتقدير الوافر لمعالي وزير التعليم المشرف العام على المركز، الذي يحث على كل ما من شأنه تثبيت الهوية اللغوية العربية، وتمتينها، وفق رؤية استشرافية محققة لتوجيهات قيادتنا الحكيمة، ويمتد الشكر لمعالي نائبه، وللسادة أعضاء مجلس الأمناء نظير الدعم والتسديد لأعمال المركز .

والدعوة موجّهة لجميع المختصين والمهتمين بتكثيف الجهود نحو النهوض بلغتنا العربية، وتحقيق وجودها السامي في مجالات الحياة .

مقدمة المحرر

تُعدّ **لسانيات المدونات** –أو المتون– (Corpus Linguistics) من العلوم الحديثة التي أحدثت تغييراً منهجياً في دراسة اللغة، حتى إن هناك من يراها منهجية تدخل في أغلب العلوم اللغوية. وعلى الرغم من ذلك فقد كانت محل جدل واسع في سبعينات القرن الماضي؛ لكن هذا الجدل خبا وصار جزءاً من الماضي.

إن الاستعانة بلسانيات المدونات في **اللغة العربية** درساً وبحثاً وتعلماً وتعليماً أمر ضروري إذا أردنا أن تأخذ لغتنا مكانها المستحق بين اللغات الحية؛ ولهذا فقد خرج هذا الكتاب ليقدّم لقراء العربية مادة حديثة من الأهمية بمكان أن يعرفوا عنها الكثير.

جاء هذا الكتاب في خمسة مباحث تشرح وتناقش القضايا الرئيسة في هذا العلم المهم. فقد قدّم لنا **محمود إسماعيل صالح** (محمود فيما بعد) في **المبحث الأول** القضايا الرئيسة في المدونات اللغوية وكيفية الإفادة منها بإلقاء الضوء على تعريف المدونات اللغوية مرورًا بلمحة عن طرق الإفادة منها وعن أنواع المدونات اللغوية ومواصفاتها وإنشائها، ومتطلبات التعامل معها، ثم مناقشة تفصيلية لمجالات الإفادة من المدونات اللغوية في فروع البحث اللساني العام والتطبيقي والحاسوبي مع التمثيل لكل فرع.

وفي **المبحث الثاني** يطوف بنا **عبدالله بن يحيى الفيفي** (الفيفي فيما بعد) في ميدان مدونات المتعلمين التي أضحت إحدى ركائز البحث في مجال اكتساب اللغة وتعليمها، مستعرضاً بعض الأسس النظرية مثل تعريف مدونات المتعلمين، وبيان أنواعها، وما الذي يميزها عن غيرها، مع استعراض المدونات العربية منها، ليتحدث بعد ذلك عن المجالات التي تمثّل أهم النماذج التي استفادت – ولا زالت تستفيد – من مدونات المتعلمين، وهي: أبحاث اكتساب اللغة وتعليمها،

والتحليل التقابلي للغة المرحلية، وتحليل الأخطاء بمساعدة الحاسب، ودراسة التطور اللغوي لدى الطلاب، وتأليف المعاجم، وتدريس اللغة وتصميم المواد التعليمية، وأخيراً البحث في معالجة اللغة الطبيعية. وقد ختم حديثه بوسم الأخطاء في اللغة العربية، مع الوقوف على بعض النماذج العملية لطريقة وسم الأخطاء.

ويأتي بعد ذلك **المبحث الثالث** الذي يتحدث فيه **عبدالمحسن بن عبيد الثبيتي** (الثبيتي فيما بعد) عن تصميم المدونات وبنائها، ويقدّم فيه إطاراً عملياً للإجراءات التي تمكّن الباحث من تحقيق الشروط العلمية الرئيسة لتصميم المدونات (التوازن والتمثيل، والمنهج العلمي، والثبات) من خلال اتباع معايير واضحة لتصميم وجمع نصوص المدونة. كما وضّح أهم ما يجب النظر إليه عند تصميم المدونات مثل لغة المدونة، وطبيعة نصوصها، وتاريخ النصوص، ومكان صدورها، والأوعية التي ظهرت فيها، والمجالات المختلفة التي تغطيها. كما تطرق إلى حقوق الملكية الفكرية، وتحديد مصادر جمع النصوص، وكيفية جمع النصوص مع مناقشة بعض المسائل التقنية عند الانتهاء من الجمع وهي ترميز ملفات نصوص المدونة وتسميتها وحفظها وإضافة معلومات مساعدة في تعزيز الفائدة من المدونة من خلال تحشيتها بمعلومات إضافية مختلفة. وتحدّث بعد ذلك عن أهم الأدوات التي تساعد في تحليل المدونات ودراستها، وهي: الإحصاءات العامة عن المدونة بمجملها، وقوائم التكرار، والكلمات المميزة للمدونات، والكشاف السياقي، والتصاحب اللفظي، ثم ختم المبحث بمثال تطبيقي لما سبق شرحه.

وبانتقالنا إلى **المبحث الرابع** يخوض بنا **عقيل بن حامد الشمري** (الشمري فيما بعد) **والثبيتي** غمار التطبيق للسانيات المدونات؛ إذ يقدمان نماذج تطبيقية، وعينات توضح طرق استخدام «لسانيات المدونات»، وسبل استثمار إمكاناتها الحاسوبية في رصد وتحليل الظواهر اللغوية، معتمدَين على «العربية

الفصحى المعاصرة» المستخدَمة في لغة الصحافة المكتوبة في كافة الأقطار العربية. وقد وقع اختيارهما على هذا الموضوع لربط لسانيات المدونات بوصفها منهجية في الدراسة والتحليل بالإشكال النظري المتعلق ببعض ظواهر العربية المعاصرة لكي يتسنى لهما مناقشة ما يتعلق بذلك من قضايا ومسائل مع التمثيل المناسب المعتمد على «لسانيات المدونات».

أما **سلطان بن ناصر المجيول** (المجيول فيما بعد) فيأتي حديثه في **المبحث الخامس** عن واقع البحث اللغوي العربي الحاسوبي، معرِّفاً بطرائق البحث اللغوي العربي الحاسوبي بصورة أولية تَرسم تمرحلات التحليل بواسطة المعالجة الحاسوبية بشكل مبسط؛ بدءًا بالتعريف بأنواع المدونات الحاسوبية للغة الطبيعية واتجاهاتها الممكنة والمحتملة والمأمولة، مرورًا بخصائص كل طريقة ومناهجها البحثية من حيث الغرض والتحليل والمعالجة، وصولًا إلى اقتراح بعض الموضوعات البحثية المهمة في هذا المجال.

ويختم الكتاب **بنظرة استشرافية** يقدّم فيها **صالح بن فهد العصيمي** (العصيمي فيما بعد) ما يمكن لباحث في هذا الميدان أن يؤمّله للغته التي يرى ضرورة أخذها مكانها بين اللغات الحية بمواكبة التطورات اللسانية والتقنية في هذا المجال.

إن الهدف من تأليف هذا الكتاب هو التعريف بلسانيات المدونات لقراء العربية وكيفية الإفادة منها في بحوثهم ودراساتهم، بالإضافة إلى استعراض الدراسات التي أُجريت على العربية في هذا الميدان. فهذا الكتاب يستعرض إمكانيات لسانيات المدونات وواقع البحث اللغوي العربي فيها والآفاق التي يمكن أن تطرقها لغتنا فيها. ويعد هذا الكتاب الأول من نوعه بوصفه مرجعاً متخصصاً في مجال المدونات اللغوية باللغة العربية، حيث إن أغلب ما هو موجود عبارة عن أبحاث ومقالات -مع أهميتها- تحدثت عن المدونات بشكل مختصر أو عن جزئيات متفرقة ضمن موضوعات أخرى؛ بينما عالج هذا

الكتاب موضوع المدونات بشكل أقرب إلى الشمولية، إضافة إلى ما أورده من التطبيقات اللغوية والحاسوبية للمدونات، والتي يمكن أن تكون أساساً لكثير من الأبحاث المستقبلية في هذا الميدان.

وهذا الكتاب يُوجَّه للباحث في العربية ولمعلمها ولدارسها. وقد كان أسلوب الكتاب ميسّراً ليتمكن من فهمه طلاب الدراسات العليا بالإضافة إلى طلاب المرحلة الجامعية؛ بل حتى الأساتذة ذوو الخبرة الطويلة في اللغة العربية يمكنهم الإفادة من كثير من المسائل المطروحة هنا إذا رغبوا في الاطلاع على ما تمّ في هذا الحقل المعرفي المهم.

ويمكن للقارئ أن **يستخدم هذا الكتاب** بالبداية من أوله إلى آخره إن لم يكن على اطلاع كافٍ بلسانيات المتون، كما يمكن لمن لديه خبرة في التخصص أن يقرأ المبحث الذي يناسبه، فالمباحث في هذا الكتاب وإن كان بينها تماسك عضوي فكل مبحث يقدم مجموعة من القضايا مستقلة بذاتها. والكتاب في الوقت نفسه يصلح مرجعاً للسانيات المدونات العربية.

إن تقديم الكتاب بهذه الصورة لم يخلُ من بعض **الصعوبات والتحديات** التي حاولنا التعامل معها بما يمكن لنا أن نقوم به، فظهر بعض التداخل في بعض مباحث الكتاب لكن كل مبحث عالج القضايا الخاصة من زاويته حتى وإن بدت متشابهة مع المباحث الأخرى. كما كان ضرب الأمثلة باللغة العربية ليس عملاً يسيراً باعتبار الفلسفيات التي يعتمد عليها هذا التخصص واعتمادها على اللغة الإنجليزية بالدرجة الأولى. أما قضية المصطلحات وترجمتها فقد شغلت الباحثين طيلة فترة تأليف هذا الكتاب، وذلك لإيجاد مصطلح ينقل المعنى الأصلي بظلاله ويعطي الدلالة العربية.

ولا يسعنا إلا أن **نشكر** مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية على هذه المبادرة الرائعة التي ستثري الدراسات العربية في هذا

الجانب، كما نخص بالشكر القائمين على سلسلة مركز الملك عبدالله بن عبدالعزيز للسانيات العربية. كما أتقدم بالشكر الجزيل للباحثين المشاركين في هذا الكتاب على إسهامهم الذي سيخدم العربية وأبناءها ومتعلميها. والشكر موصول للذين يسألون ويستفسرون بشكل دائم عن لسانيات المدونات من طلاب وباحثين والذين كانوا الوقود الدافع للتفكير في هذا الكتاب وتأليفه.

المحرر/ صالح بن فهد العصيمي
الرياض
٢٦ صفر ١٤٣٦هـ
salehosaimi@yahoo.com

تعريف بالباحثين المشاركين في التأليف

سلطان بن ناصر المجيول: أستاذ علم لغة المدونات والتحليل اللغوي الآلي المساعد في جامعة الملك سعود، الرياض. درس الدكتوراه في مركز الدراسات الشرقية الأوسطية ومركز أبحاث المعجم في جامعة إكسيتر، بريطانيا. عمل في تعليم اللغة العربية لغير الناطقين بها منذ عام ٢٠٠٤م حتى عام ٢٠٠٩م، كما عمل محررًا -ومازال- للمجلة العالمية للدراسات الشرق الأوسطية International Journal of Middle East Studies، وعمل محررًا لأبحاث إنجليزية ذات علاقة بموضوعات إثنوجرافية وأنثروبولوجية ولغوية اجتماعية في الخليج العربي بمركز الدراسات الخليجية في جامعة إكسيتر، وعمل محاضرًا متعاونا مدة عام ونصف في جامعة إكسيتر بتدريس مادة القراءة بالعربية للمتقدمين والنقل الترجمي لنصوص القرآن الكريم لطلاب السنة الرابعة لمرحلة البكالوريس. ورأس قسم التطوير في كلية الآداب بجامعة الملك سعود مدة عام واحد.

صالح بن فهد العصيمي: أستاذ الدراسات العليا المساعد للغويات التطبيقية والتربية في جامعة الإمام محمد بن سعود الإسلامية، الرياض. درس الدكتوراه في كلية التربية في جامعة ليدز، بريطانيا. وقد عمل مستشاراً ومحكّماً لمايكروسوفت في عدة مشاريع برمجية خاصة باللغة العربية، كما عمل مستشاراً في مكتب التربية في مقاطعة ليدز وفي المجلس البلدي في مدينة ليدز وعضواً في هيئة الاستئناف في الدعاوى المرفوعة ضد مكتب التربية، وممثلاً للمسلمين في المجلس الاستشاري الدائم للتربية الدينية (SACRE: Standing Advisory Council of Religious Education). وهو معار حالياً لشركة تطوير للخدمات التعليمية مديراً لمشروع تطوير تعليم اللغة العربية في مشروع الملك عبدالله بن عبدالعزيز لتطوير التعليم العام. وقد حكّم العديد من البحوث والمشاريع

المتلعقة بلسانيات المدونات لجهات علمية وأكاديمية، كما درّب معلمي وأساتذة اللغة العربية لغة ثانية في إندونيسيا والصين والمملكة العربية السعودية.

عبدالله بن يحيى الفيفي: معهد تعليم اللغة العربية، جامعة الإمام محمد بن سعود الإسلامية، الرياض. متخصص في معالجة اللغة الطبيعية. شارك في عدد من المشاريع العلمية في مجال معالجة اللغة العربية حاسوبياً، له عدة أبحاث منشورة حول مدونات المتعلمين، إضافة إلى مشاركته في تحكيم عدد من الأبحاث العلمية المتعلقة بمعالجة اللغة الطبيعية لدى عدد من الدوريات العلمية والمؤتمرات. قام بإنشاء المدونة اللغوية لمتعلمي اللغة العربية، وبناء بعض التطبيقات الحاسوبية للمساعدة على وسم وتصحيح الأخطاء في النصوص العربية بمساعدة الحاسب. مهتم بعلم اللغة الحاسوبي، وتعليم اللغة بمساعدة الحاسب، إضافة إلى تعلم وتعليم اللغة العربية باعتبارها لغة ثانية.

عبدالمحسن بن عبيد الثبيتي: أستاذ البحث المساعد في المركز الوطني لتقنية الحاسب والرياضيات التطبيقية بمدينة الملك عبدالعزيز للعلوم والتقنية وهو الباحث الرئيس لمشروع المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية. عمل في مجال إدارة تقنية المعلومات لعشرين عاماً قبل انتقاله للعمل في مدينة الملك عبدالعزيز للعلوم والتقنية. حاصل على درجة الدكتوراه من جامعة ساري ببريطانيا من قسم الحوسبة. نشر العديد من الأبحاث في المؤتمرات والدوريات المحكمة. مهتم بالمصادر اللغوية العربية ومعالجة اللغة العربية والتنقيب في النصوص والتعقب الآلي لتطور المعرفة في المجالات المتخصصة.

عقيل بن حامد الشمري: أستاذ اللسانيات التطبيقية المساعد، ورئيس قسم تدريب المعلمين بمعهد اللغويات العربية، بجامعة الملك سعود، الرياض. حصل على درجتي الماجستير والدكتوراه في جامعة كوينزلاند، أستراليا. اهتماماته البحثية تشتمل على اكتساب اللغة والحرافة/literacy، وأبحاث الكتابة باللغة

العربية، وتعلم اللغة بمساعدة الحاسوب. عمل محاضراً متعاوناً في تدريس مادة «اكتساب اللغة الثانية» لطلاب الماجستير في جامعة كوينزلاند، بالإضافة إلى تدريس مادة «الحاسوب واللغة» لطلاب الماجستير في معهد اللغويات العربية. يعمل في مجال تعليم اللغة العربية لغير الناطقين بها، وشارك في العديد من الدورات التدريبية لمعلمي العربية في عدد من الدول المختلفة. كما يعمل مستشاراً غير متفرغ في كل من مشروع الملك عبدالله بن عبدالعزيز لتطوير التعليم العام (مشروع اللغة العربية)، ولجنة التخطيط اللغوي والسياسة اللغوية في مركز الملك عبدالله لخدمة اللغة العربية.

محمود إسماعيل صالح (الصيني سابقا): تخرج في جامعة جورجتاون بواشنطن العاصمة في ١٩٧٢ (تخصص لسانيات تطبيقية)، وهو أستاذ اللسانيات التطبيقية في جامعة الملك سعود. وقد أسس أول معهد متخصص لتعليم العربية للناطقين بغيرها في عام ١٩٧٥ في الجامعة، كما أسس البنك السعودي للمصطلحات العلمية (باسم) في مدينة الملك عبدالعزيز للعلوم والتقنية في العام ١٩٨٣، وأنشأ مركز الترجمة في جامعة الملك سعود في العام ١٩٨٩. بالإضافة إلى عمله الأكاديمي في قسم اللغة الإنجليزية في كلية الآداب بجامعة الملك سعود، يعمل خبيراً في هيئة الخبراء بمجلس الوزراء السعودي (مجال الترجمة الرسمية) وعضواً باللجنة الاستشارية بمركز الملك عبدالله الدولي لخدمة اللغة العربية، ومستشاراً غير متفرغ في مشروع الملك عبدالله لتطوير التعليم العام (مجال تطوير معلمي اللغة العربية). تشمل أعماله أكثر من أربعين دراسة في اللسانيات والترجمة وتعليم اللغات وصناعة المعاجم والمصطلحية. ونشر بمفرده أو بالتعاون مع آخرين أكثر من خمسة معاجم أحادية اللغة وثنائيتها وستين كتاباً في تعليم العربية للناطقين بغيرها.

المبحث الأول

المدونات اللغوية وكيفية الإفادة منها

محمود إسماعيل صالح

المقدمة

من الأمور التي لا يجهلها أحد أن الحاسوب هو أداة القرن الحالي دون منازع. فقد دخل تقريبا كل بيت وكل مؤسسة عامة وخاصة، وأدى استخدامه في مجالات الحياة العلمية والعملية إلى تطورات كبيرة في الحقول المختلفة. ومن هذه المجالات ميدان اللغة التي تميز بها الإنسان عن سائر المخلوقات. ويمكننا الإفادة من الحاسوب في النشاط اللغوي بوجوه متعددة، يمكننا تلخيصها في ثلاث صور رئيسة هي: الاستعمال العام (مثل منسق النصوص والترجمة بمعاونة الحاسوب وقواعد بيانات المصطلحات)، والاستعمال الخاص (أداة للبحث اللساني، كما هو الحال مع لسانيات المدونات اللغوية واستعمال برمجيات قواعد البيانات في تحليل الأخطاء اللغوية)، والبحث اللساني الحاسوبي المتخصص (مثل تطوير البرمجيات المختلفة للتعامل مع اللغة، مثل برمجيات التعرف على الأصوات وتوليفها والترجمة الآلية).

غير أن أحدث وأهم مجال لعبه الحاسوب في خدمة البحث اللغوي هو مجال المدونات اللغوية corpora، ومفرده (مدونة corpus)، وهو موضوع هذا الكتاب. ولعل مما يلفت انتباه الباحث العربي أنه على الرغم من مرور نصف قرن تقريبا على أول مدونة إلكترونية في اللغة الإنجليزية وحوالي ثلاثة عقود على البحث اللساني المبني على المدونات المحوسبة فإننا نلحظ قلة إلمام اللسانيين العرب بها وبإمكاناتها غير المحدودة. من هنا نأمل أن تكون هذه الدراسة بمثابة حافز للعلماء العرب للخوض في غمار هذا المجال وفي الإفادة من المدونات اللغوية العربية في بحوثهم اللغوية.

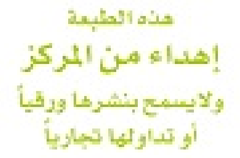
أولاً: تعريف المدونة اللغوية

لعل أبسط تعريف للمدونة اللغوية هو: مجموعة من النصوص اللغوية الشفوية أوالمكتوبة الموثقة (من حيث المصدر والتاريخ والنوع كحد أدنى).

يقول ماكنري وزميلاه (McEnery et al.، ٢٠٠٦:٤) «في اللسانيات الحديثة، المدونة اللغوية يمكن تعريفها بأنها مجموعة من نصوص اللغة الطبيعية»، ولكن بشكل أدق «يجب أن نضيف أن المدونات الحاسوبية نادرًا ما تكون مجموعات اعتباطية من النصوص. فهي تجمع بصورة عامة لغايات محددة وغالبا ما تجمع لتكون (بشكل غير رسمي) ممثلة لبعض اللغة أو نوع من النصوص» (Leech، ٢٩٩٢:١١٦).

ومن التسميات الشائعة -إلى حد ما- اسم «الذخيرة اللغوية»، والتي أطلقها رائد العمل العربي في مجال المدونات المحوسبة اللساني عبدالرحمن حاج صالح، غير أن تعريفه للذخيرة اللغوية كما ورد في دراسته التي نورد مقتبساً منها لاحقاً يختلف قليلاً عن المدونة اللغوية بالتعريف الذي ذكرناه (انظر عبدالرحمن صالح، ١٩٩٩). كما أن مها الربيعة (١٤٣٣هـ) تطلق عليها «ذخيرة نصية» كما هو واضح من عنوان مشروعها المسمى «الذخيرة النصية الفصحى لجامعة الملك سعود». وهناك من أسماها بالمكنز، كما فعل عبدالغني أبو العزم (حيث يتحدث عن «مكنز صخر» في مقاله «اللغة العربية والمعالجة الآلية: برامج صخر نموذجاً»). كذلك نجد دراسة لسلوى حمادة أطلقت فيها على المدونات اللغوية «المدونات النصية» (سلوى حمادة، ٢٠١٤) غير أنها أسمتها أيضاً «المدونات اللغوية» في دراسة أخرى (سلوى حمادة، ٢٠١١). وفي دراسة حديثة لصالح العصيمي (٢٠١٣) أطلق «لسانيات المتون» على لسانيات المدونات اللغوية.

ولكن يبدو أن مدونة (لغوية) قد تكون الأكثر شيوعاً، وهو الاسم الذي أطلق على أشهر مدونتين عربيتين معروفتين (مدونة مدينة الملك عبدالعزيز للعلوم والتقنية في الرياض، ومدونة مكتبة الإسكندرية التي سنتحدث عنهما في صفحات لاحقة من هذا المبحث). وقد يتحرج البعض من استعمال كلمة مدونة لأنها تستخدم مقابل كلمة blog أي المدونة الشخصية في الشابكة (الإنترنت)؛



لذلك كان الاقتراح بإضافة اللغوية للتمييز بينهما. (نود الإشارة بأننا قد نقتصر على مصطلح مدونة فقط، بدلاً من مدونة لغوية، من باب الاختصار ولوضوح المعنى في سياق هذا العمل).

من الناحية النظرية، قد تكون المدونة مجموعة من النصوص المجموعة يدويّاً، كما نجد في المدونات السابقة لعصر الحاسوب، أو حاسوبيّاً، كما هو الغالب اليوم.

لكن نود أن ننبه على أن «المدونة» في سياق لسانيات المدونات اللغوية يقصد بها المدونة اللغوية المحوسبة، أي المخزنة رقمياً في الحاسوب. لذلك نجد أن البعض يتحدثون عن لسانيات المدونات باسم لسانيات المدونات الإلكترونية electronic corpus linguistics.

وجدير بالذكر في هذا المجال أن الشبكة العنكبوتية (الإنترنت) أو الشابكة يمكن اعتبارها مدونة لغوية شاملة، كما يشير الباحثان كيلغاريف وغرينفنستيت Adam Kilgarriff and Greogroy Grefenstette في بحثهما المعنون «الشبكة بوصفها مدونة لغوية Web as Corpus» حيث يقولان في مقدمة بحثهما أن الشابكة «تشتمل على مئات البلايين من الكلمات من النصوص ويمكن استعمالها لجميع أنواع البحث اللغوي» (Adam Kilgarriff and Greogroy Grefenstette، ٢٠٠٣).

ثانيًا: طرق الإفادة من المدونات اللغوية

من متابعة الأعمال البحثية المختلفة، لا بد لنا من التنويه إلى أن هناك اختلافاً بين نوع المدونة (يدوية أو محوسبة) من جهة وبين طريقة الإفادة منها في البحث اللغوي. فحتى المدونة المحوسبة يمكن البحث فيها يدوياً أو بأسلوب شبه يدوي، وهو ما نعتقد أنه حدث مع مدونة جريدة الجزيرة وعدد من الأبحاث التي أجريت عليها (كما في: دراسات في علم اللغة النصي...، ٢٠١٢) . وقد يتم

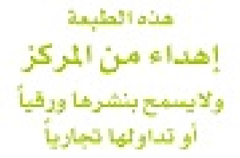
إجراء البحث آلياً، كما هو الشأن في معظم الدراسات التي تمت على المدونات الإنجليزية، بدءاً بمدونة براون والدراسات الأخرى الواردة في بايبر وزملائه (Biber، Conrad and Reppen، ١٩٩٨) وكذلك ماكنري وزميليه (McEnry، Tony، Xiao، Richard and Tono، Yukio، ٢٠٠٦). وأخيراً المشروع الذي أنجزه الباحثان الأمريكيان تم بَكوولتر وديلويرث باركنسون في معجم الألفاظ الشائعة في اللغة العربية (Buckwalter، Tim and Parkinson، Dilworth، ٢٠١١)، وكذلك أطروحة الدكتوراة لرجب الزهراني (انظر أدناه). (لمزيد من المعلومات في هذا المجال، انظر «أعمال مبنية على مدونات لغوية عربية» المنشورة في: مدونة الدكتور محمود إسماعيل صالح على الشابكة).

ولكن ينبغي لنا أن نفرق بين «الدراسات المبنية على المدونات corpus-based ، وهي التي تفيد من المدونات اللغوية في البحث اللساني وفق إطار محدد مسبقاً، كالتعرف على شيوع كلمات أو تراكيب معيّنة أو التثبُّت من ظواهر لغوية معيّنة، والدراسات التي توجهها المدونات corpus- directed، والتي تنتج من النظر في المدونات اللغوية، حيث يصل الباحث إلى حقائق جديدة لم يكن يبحث عنها بالضرورة. لذلك لا يفرق البحث الموجّه بالمدونات اللغوية بين المعجم والنحو والتداولية وعلم الدلالة والخطاب، بل تتبع منهجاً كليّاً في البحث اللساني» (McEnery et al.، ٢٠٠٦:١٠-١١، للمزيد من مناقشة الفروق بين النوعين من الدراسات، انظر ماكنري وزميليه، ٢٠٠٦: ٨-١١). ولكن الفرعين كليهما يقعان تحت مسمى لسانيات المدونات اللغوية corpus linguistics.

ثالثاً: أنواع المدونات اللغوية

يمكننا تصنيف المدونات اللغوية من زوايا مختلفة، كالعموم والخصوص والمعاصر والتاريخي، ومن حيث الوسم و عدمه ...

أ- من حيث العموم:



١- المدونة العامة:

هي مدونة أنشئت لأغراض مختلفة ومن نصوص متباينة. ومن أمثلتها: «مدونة اللغة العربية» و«المدونة الوطنية البريطانية British National Corpus».

٢- المدونة الخاصة بنوع من النصوص (نص ديني، شعر، نثر، قصة، كتابة تقنية أو علمية...)، أوخاصة بمنطقة جغرافية (بلد أو منطقة معينة)، أوخاصة بمؤلف ما، أوخاصة بنص ما «مدونة القرآن الكريم Qur'an Corpus».

٣- المدونة المقارنة أو المتشابهة comparable:

هي المدونة التي تشتمل على نصوص متشابهة من حيث المحتوى (مثلاً رسائل جامعية أو مقالات علمية في مجال ما) بلغتين أو صورتين مختلفتين للغة نفسها. ومن أمثلة ذلك المدونات التي تهدف دراستها للمقارنة بين كُتّاب مختلفين في كل من اللغتين أو نوعية اللغة، مثل أساليب أهل اللغة ومتعلّميها من غير أهلها. وهذا النوع من المدونات هو السائد في الدراسات اللغوية التقابلية constrastive.

٤- المدونة المتوازية parallel:

تشتمل على مجموعة من النصوص المتماثلة بلغتين مختلفتين (مثلاً أحد النصين ترجمة للنص من لغة أخرى). ويستفاد من هذا النوع من المدونات في دراسات الترجمة والمترجمين.

وجدير بالذكر أن مشروعات وبرمجيات الترجمة الآلية المبنية على أسس إحصائية (مثل برمجيات شركة آي بي إم IBM) تعتمد بشكل أساس على هذا النوع من المدونات. كما أن هناك برنامجاً

على الشابكة لتطوير برمجيات الترجمة الآلية المبنية على المدونات المتوازية: Moses Statistical Machine Translation System (انظر قائمة المراجع).

٥- مدونة المتعلمين learner (متعلّمي اللغة خاصة من غير أهلها):

تتكون المدونة من أعمال أنتجها متعلمون للغة (تلقائية spontaneous، أو مستنطقة elicited). وتفيد هذه المدونات في تحليل أخطاء الدارسين من خلفيات ومستويات مختلفة، من ثم تفيد أيضا في دراسات اكتساب اللغة الثانية. ومن أمثلة هذه المدونات: مدونة كامبردج للمتعلمين Cambridge Learners' Corpus، و مدونة لونقمان للمتعلمين Longman Learners' Corpus. وللغة العربية مدونات للغة المتعلمين: إحداها التي أنشأتها الباحثتان الماليزيتان (انظر Hassan، Haslina and Nurahihan Mat Daud، ٢٠١١)، والأخرى تسمى «مدونة متعلمي العربية المكتوبة The L2 Written Arabic Corpus) التي تم إنشاؤها في جامعة أريزونا بالولايات المتحدة الأمريكية بمركز المصادر التعليمية عن الثقافة واللغة ومحو الأمية (CERCLL – Center for Educational Research in Culture, Language and Literacy). وتعتبر هذه المدونة التي تتكون من ٣٠٠ مقالة نموذجاً لما يمكن أن يكون عليه هذا النوع من المدونات، حيث يذكر الباحثون ما يلي عنها:

«تم وضع المقالات المطبوعة في قاعدة بيانات قابلة للبحث، حيث تم وسمها بمستوى المتعلم (مبتدىء ومتوسط ومتقدم)، ونوع المتعلم (دارس لغة ثانية أو التراث)، ونوع النص genre (وصف وسرد وتعليمات)» (انظر: http://l2arabiccorpus.cercll.arizona.edu). وسيتحدث المبحث الثاني بكامله عن مدونات المتعلمين بشيء من التفصيل.

٦- المدونة التعليمية pedagogic:

وتشمل المواد التعليمية لمنهج ما لأغراض تعليمية مختلفة. (انظر O'Keeffe، Anne، McCarthy، Michael and Carter، Ronald، ٢٠٠٧). ويستفاد من هذه المدونات في تعليم اللغة وفي فحص المواد التعليمية ودراستها.

ب- من حيث الزمن:

٧- مدونة تاريخية historical/diachronic:

وهي المدونة التي تشتمل على نصوص من عصور مختلفة، بصورة متوازنة، للتعرف على التطور الذي طرأ على اللغة وعلى استعمالات ألفاظها ومعانيها وعلى تراكيبها إلى غير ذلك. وهذا النوع من المدونات يجب أن يكون الأساس لأي عمل معجمي تاريخي يعتمد على الاستقراء المنهجي، بدلاً من الاجتهادات الشخصية والحدس والتخمين. ومن أحسن الأمثلة لهذا النوع من المدونات: ARCHER وتشتمل على نصوص للإنجليزية البريطانية والأمريكية للفترة ما بين ١٦٥٠ و ١٩٩٠.

٨- مدونة المتابعة monitor (أي متابعة التطور اللغوي):

نجد أن بعض مؤسسات النشر المعجمي، مثل دار كولينز البريطانية Collins، تهتم بمثل هذا النوع من المدونات التي ترصد المستحدثات المعجمية خاصة، عن طريق التحديث المنهجي للمدونة (قاعدة البيانات اللغوية)، وذلك بغرض تحديث معاجمها بين الفينة والفينة وإضافة ما يطرأ على اللغة من ألفاظ جديدة أو ملاحظة الاستعمالات الطارئة لبعض مداخل تلك المعاجم. (انظر Susan Hunson، ٢٠٠٢:١٥-١٦).

جـ- من حيث نوع اللغة:

٩- اللغة المنطوقة/المحكية، مثل نصوص هاتفية، أو تعليقات إذاعية، ومحادثات يومية. وهناك عدد محدود من هذه للهجة المصرية. ويجب التنبيه إلى الفرق بين المدونة البسيطة التي تحوّل النص المنطوق إلى نص مكتوب فحسب وبين المدونة التي تسجل بصورة دقيقة النص الشفوي (من حيث الوقفات والنبر والتنغيم، كذلك مواقع التردد والتكرار... إلى غير ذلك مما يحدث أثناء الكلام). ولكل نوع أغراضه وفوائده. (من أمثلة هذا النوع من المدونات: مدونة بيرغن لإنجليزية مراهقي لندن (كولت) Bergen Corpus of London Teenage English (COLT)، ومدونة الإنجليزية المنطوقة للمحترفين Corpus of Spoken Professional English، ومدونة كول هوم CALLHOME Corpus الخاصة بمحادثات هاتفية باللهجة المصرية (انظر قائمة المدونات).

ويفضل أن تتوفر تسجيلات النصوص الصوتية للباحثين للرجوع إليها عند الحاجة.

١٠- مدونة اللغة المكتوبة، وهذا هو الغالب على المدونات اللغوية المختلفة.

ولكن هناك مدونات تجمع بين اللغة المحكية والمكتوبة، مثل «المدونة الوطنية البريطانية BNC».

د- من حيث الوسم:

١١- مدونة خام: ويقصد بها المدونة التي تخلو من أية تحشيات annotations (صرفية أو نحوية أو معجمية أو غيرها). وينطبق هذا على معظم المدونات اللغوية المعروفة، خاصة العربية منها.

١٢- مدونة موسومة tagged: وهي المدونة التي بها تحشيات (صرفية أو نحوية أو معجمية أو غيرها)، حيث نجد وسماً للكلمات، مثلاً يبيّن قسم الكلام الذي تنتمي إليه part of speech (POS). ومن أفضل الأمثلة على هذا النوع من المدونات: Lancaster Parsed Corpus (مدونة لانكستر المعرَبَة للغة الإنجليزية). ولأمثلة للمدونات اللغوية العربية المعروفة، انظر القسم الأخير من هذا المبحث.

رابعاً: مواصفات المدونات اللغوية

من الأمور التي يجب أن تؤخذ بعين الاعتبار عند تصميم المدونات وتقويمها ما يلي:

أ- الحجم الكبير (ويحسب بعدد كلمات المدونات):

غالباً ما يكون ذلك ملايين الكلمات للمدونات العامة. ولكن ذلك قد لا ينطبق بالضرورة على المدونات الخاصة.

ب- الشمول وتمثيل استعمالات اللغة representativeness:

ويعني ذلك أن تشتمل المدونة على نصوص تمثل استعمالات اللغة المختلفة (الشفوية والمكتوبة)، في ضوء الهدف من إنشاء المدونة. فلا تقتصر المدونة اللغوية على استعمال أو أسلوب أو منطقة جغرافية معينة مثلاً.

جـ - التوازن:

من شروط المدونة الجيدة أن يكون هناك توازن بين أنواع أو فئات النصوص والتخصصات وغير ذلك مما يشمله معيار التمثيل، فلا يطغى مؤلف أو لهجة أو جنس أدبي على غيره. (انظر: Biber ،D.، ٢٠٠٣، ٨/٢٤٣:٤-٥٧).

ولأمثلة لتطبيق هذه المعايير، انظر مثلاً (Biber، ٢٠٠٣; McEnery et al، ٢٠٠٦:١٧) في وصفهم للمدونة الوطنية البريطانية British National Corpus،

وانظر أيضاً ما سنورده عن المدونة اللغوية العربية التي أعدتها مدينة الملك عبدالعزيز للعلوم والتقنية في الرياض، والمدونة اللغوية العربية الدولية التي أُنجزت في مكتبة الإسكندرية.

على سبيل المثال، نجد أن مدونة براون الرائدة اشتملت على خمسمئة «نوع/ جنس أدبي genre» مختلف من الكتابات المنقحة edited، يمثل كل منها نصاً يتكون من ٢٠٠ كلمة (المجموع: مليون كلمة من النصوص المنشورة). وفي المدونة الوطنية البريطانية تم التأكيد على أن «الكلام الذي يجمع يؤخذ من عينة تراعي التوازن من حيث العمر والجنس والطبقة الاجتماعية والمنطقة اللهجية» (Meyer، ٢٠٠٢:١٨).

ومن الدراسات المهمة التي عالجت معايير تصميم المدونات اللغوية مقالة سو أتكنز وزميليها (Sue Atkins et al.، ١٩٩٣)، ومجموعة الدراسات التي وردت في الكتاب الذي حرره مارتن وين Wynne بعنوان: تطوير المدونات اللغوية: دليل للممارسات الجيدة (Developing Linguistic Corpora: a Guide to Good Practice) الذي يجمع بين دفتيه مقالات مهمة لعدد من رواد العمل في هذا المجال، وقد صدر في العام ٢٠٠٥.

خامسًا: إنشاء المدونة اللغوية

قبل الشروع في مناقشة هذا الإجراء، نود أن نشير إلى أنه قد يتم جمع النصوص بطريقة عشوائية أو منتظمة (وفق أسس محددة). كما يتم الإفادة منها كذلك بصور مختلفة، كالبحث اليدوي، كما فعل اللغويون العرب الأوائل عند وضع قواعد اللغة العربية ومعاجمها، ودليل ذلك الاستشهادات التي نجدها في كتب النحو والمعاجم. (انظر أحمد مختار عمر «مصادر اللغويين العرب» في كتابه (البحث اللغوي عند العرب، ١٩٨٨). كذلك نجد مثالاً على ذلك ما قام به الباحثون في علم اللغة التاريخي في أوربا وما قام به اللسانيون البنيويون من

أمثال بلومفيد وسابير في دراساتهم للغات الهنود الحمر وغيرهم ممن عملوا في اللسانيات الميدانية، وكذلك ما قام به لانداو وفاخر عاقل وداود عبده في دراساتهم الإحصائية للكلمات العربية في نصوص مختلفة (انظر داود عبده، ١٣٩٩هـ)، وما قام به مايكل وست Michael West في بريطانيا ومن قبله ومن جاء بعده من الباحثين في دراساتهم المعجمية الإحصائية المبنية على مدونات يدوية/مكتوبة.

أما إذا نظرنا إلى المدونات اللغوية من المنظور اللساني الحديث، فنود الإشارة إلى أن من الأمور التي يجب أخذها بعين الاعتبار عند تصميم المدونة اللغوية وإنشائها ما يلي:

أ- تحديد الهدف من المدونة، مثلًا دراسة ظواهر لغوية معيّنة أو التعرُّف على الملامح المميّزة لنوع من النصوص، أو لأسلوب كاتب معيَّن، أو التطور اللغوي وغير ذلك. (المبحث الثالث بكامله من هذا الكتاب مكرّس لتصميم المدونات وبنائها).

ب- تحديد النصوص المناسبة، من حيث النوع والزمان والمكان والمصدر. فإذا كانت الغاية إجراء دراسات على اللغة العربية المعاصرة في صورتها المكتوبة، فلا بد أن تشمل المدونة نصوصًا تمثل أنواع النصوص والأجناس الأدبية المختلفة من أدب (شعر، ونثر قصصي وغير قصصي، ومسرح، ومقالات أدبية...)، والتخصصات العلمية المختلفة وكتابات لمؤلفين مختلفين من سائر الدول العربية، وأعمالاً أنتجت في الفترة المحددة، وهكذا.

جـ- تحديد نسبة كل نوع، وفق أسس علمية ومنهجية: ويخضع هذا لمعايير إحصائية (مثل نسبة الكتابات الصحفية والكتب العلمية المنتجة في

فترة محددة إلى الأعمال الأخرى) ، كما يستفاد في ذلك من آراء الخبراء والمتخصصين في مجالات المعرفة المختلفة.

د- البحث عن مصادر النصوص (مكتوبة، مطبوعة، رقمية، صوتية أو سمعية-بصرية).

هـ- جمع النصوص من مصادرها المختلفة.

و- تحويلها إلى نصوص رقمية موحدة من حيث التشفير والترميز.

ز- وسم tagging النصوص (العنونة الصرفية والنحوية – كما يسميها صوالحة (٢٠١١) – أي تحديد قسم الكلام لكل لفظة مثلاً) إذا أمكن. جدير بالذكر أن بعض الباحثين يستخدم مصطلح التحشية annotation بدلاً من الوسم. ويشرح ليتش Leech (٢٠٠٥) التحشية بقوله: «تحشية النص هي إضافة معلومات تفسيرية للمدونة. مثلًا، من الممارسات الشائعة للتحشية إضافة الوسوم tags أو العناوين labels لبيان أقسام الكلام التي تنتمي إليها كلمات النص». وله دراسة مستفيضة في هذا الموضوع يناقشه من زوايا متعددة (انظر Leech، ٢٠٠٤).

وفي جميع الأحوال لا بد من توفر حد أدنى للتعليقات أو التعليم markup على كل نص (المصدر، التاريخ ، المؤلف مثلاً). وتقدم لنا لطيفة السليطي وإريك أتويل دراسة مفيدة في تصميم مدونة للغة العربية المعاصرة (Al-Sulaiti, L and Atwell، E، ٢٠٠٦).

سادسًا: وسائل جمع نصوص المدونة وتخزينها

من المعروف أن أي مدونة لغوية لا بد فيها من جمع نصوصها من مصادر مختلفة وتخزينها وفق معايير محددة، تتفق مع الأهداف التي تُنشأ المدونة من

أجلها. ويختلف هذا الجمع تبعاً لنوع النص وطبيعته ومصدره. ونورد فيما يلي الأساليب المختلفة لهذا الجمع.

أ- النصوص الشفوية:

هناك طرائق عدة لتخزين النصوص الشفوية حاسوبيا: (١) تحويل النص إلى نص مكتوب باستخدام الإملاء التقليدي. ويمكننا استخدام برامج الإملاء الآلي لهذا الغرض، مع ضرورة المراجعة الدقيقة لها. (٢) كتابة النص كتابة صوتية، أي باستخدام رموز بديلة للألفباء الصوتية phonetic alphabet. وهذا مطلوب في تخزين النصوص العامية أو اللهجية وكذلك لغة الطفل. (٣) الكتابة الصوتية مع إضافة رموز للدلالة على الوقفة القصيرة والطويلة والتردد وما شابه ذلك من ظواهر يتسم بها الكلام المنطوق والحوارات. وفي جميع الأحوال لا بد لنا من تحديد المتحدث في الحوار. (لمزيد من التفاصيل بخصوص كتابة النصوص الشفوية، انظر Thomson، P.، ٢٠٠٤).

ب- النصوص المكتوبة يدوياً:

الوسيلة المعروفة هي تخزين النصوص يدوياً عن طريق لوحة المفاتيح. ولكن يمكننا تخزين هذه النصوص نظرياً على الأقل عن طريق الإملاء الآلي بقراءة النصوص واستعمال نظام تحويل النص الشفوي إلى نص مكتوب (كلام إلى نص speech-to-text). وقد ظهرت في الآونة الأخيرة تطبيقات تحوّل الخط اليدوي إلى مادة مطبوعة رقمية.

جـ- النصوص المطبوعة:

هناك وسيلتان لتخزين هذه النصوص حاسوبياً: الأسلوب التقليدي وهو استخدام لوحة الطباعة/المفاتيح في الحاسوب. أما الأسلوب الآخر فهو باستخدام قارئة المحارف البصرية (optical character

reader: OCR)، حيث يقوم البرنامج بمسح النص المطبوع ثم تحويله إلى نص رقمي. غير أن هناك مشكلات تواجه هذه الوسيلة، منها: (١) محدودية برامج قراءة المحارف البصرية من حيث تعاملها مع أنماط معينة (أي أشكال محدودة من الحروف)، (٢) ضرورة الوضوح التام لحروف النص المطبوع، (٣) وجود نسبة تقل أو تكثر من الخطأ في عملية قراءة المحارف، مما يتطلب تدقيقاً يدوياً. (من أمثلة هذا النوع من البرمجيات «القارئ الآلي» من إنتاج شركة صخر المعروفة، والذي يتعامل مع الحروف العربية واللاتينية).

د- النصوص الرقمية:

هناك عدة أنواع من النصوص المتاحة رقمياً، لعل أهمها ما يلي:

١- النصوص الناتجة عن استخدام برمجيات معالجة النصوص/الكلمات word processors.

٢- النصوص المتاحة على الشابكة (الإنترنت).

٣- النصوص المتاحة في صورة PDF.

٤- المواد المتاحة لدى شركات الطباعة والناشرين.

وتحتاج هذه النصوص كلها إلى: (١) تحويلها إلى نصوص خالية من المعلومات الخاصة بالشكل format، أي نصوص رقمية بسيطة plain text. (٢) مراعاة توحيد الشفرة code المستخدمة، لكي تتعامل معها برامج تحليل النصوص مثل المكشاف السياقي concordance. (٣) تحويل النص المتاح في صورة PDF إلى نص من الحروف قابل للتعديل عليه، من خلال ما يسمى بمحوّل البي دي إف PDF converter، ثم إخضاعه للخطوتين الأولى والثانية السابقتين.

سابعًا: مصطلحات مهمة في مجال لسانيات المدونات

سنورد بعضاً من أهم مصطلحات هذا العلم. ولمزيد من المصطلحات انظر «مصطلحات لسانيات المدونات اللغوية» في مدونة محمود إسماعيل صالح على الشابكة (الإنترنت). كما سيرد مسرد مصطلحات في نهاية هذا الكتاب.

أ- تعليم المدونة markup: وهو إعطاء معلومات عن مصدر المدونة وتاريخ النص ومؤلف النص ونوع النص genre ويشمل الجنس الأدبي وغيره. ويكون ذلك خارج النص. لكن يلاحظ أن بعض الباحثين يستعملون المصطلح ليشمل جميع أنواع المعلومات التي تضاف إلى النص سواء داخله أو خارجه، مما يعنى اشتماله على التحشية والوسم.

ب- التحشية annotation: وهي إضافة معلومات في داخل النص، مثل اا‌لوسم tagging. وهناك أنواع مختلفة للوسم، منها:

(١) قسم الكلمة (اسم، فعل، حرف). ويجدر ملاحظة أن التقسيم العربي التقليدي يحتاج إلى مراجعة وتفصيل أكثر عند وسم الكلمات، حتى يكون التحليل أكثر فائدة ودقة، بناء على أهداف المدونات التي تختلف عن أهداف واضعي النحو وتقسيماته. (انظر محمود إسماعيل الصيني، ١٤١٢) «نحو معجم عربي للتطبيقات الحاسوبية»، وانظر تطبيقاً عملياً لنوع جديد من التقسيم في المعجم الذي أعده بكوولتر وباركنسون (Buckwalter and Parkinson، ٢٠١١)، وهما من رواد العمل في مجال مدونات اللغة العربية في الولايات المتحدة الأمريكية.

ولا بد من الإشارة إلى أن كتب النحو العربي تزخر بمصطلحات تتعلق بتصنيف الكلمات العربية يمكننا الإفادة منها، من مثل: ضمير، اسم موصول، اسم جامد، اسم مشتق (اسم فاعل، اسم

مفعول، صيغة مبالغة، صفة مشبهة، اسم آلة، اسم مكان)، اسم جمع، علم، مصدر، ظرف (مكان، زمان)، فعل (لازم، متعد لمفعول، متعد لمفعولين، ناقص، ماض، مضارع، أمر...)، حرف (جر، نصب، جزم، عطف، نفي...). ويجب وسم الكلمات بمثل هذه التصنيفات لإجراء الدراسات المعجمية والنحوية. (انظر مجدي صوالحة، ٢٠١١؛ وأيضاً Kais Dukes and Nizar Habash، ٢٠١٠، وكذلك مقال صوالحة وأتويل، ٢٠١١ Sawalha and Atwell).

(٢) الوسم النحوي أو الإعرابي (لبيان الوظيفة النحوية/الحالة الإعرابية):

مثالان للوسم الإعرابي:

الجملة الأولى: «الإيمان نور القلوب»

الوسم الإعرابي: (مب (الإيمان) خب (مض (نور) مض إل (القلوب)))

الشرح:

الإيمان = مبتدأ (مب)

نور القلوب = خبر (خب)

نور= مضاف (مض)

القلوب = مضاف إليه (مض إل)

المستوى الأول: الإيمان (مبتدأ) + نور القلوب (خبر)

المستوى الثاني: نور (مضاف) + القلوب (مضاف إليه)

(يلاحظ أن عدد الأقواس يشير إلى مستوى التحليل)

الجملة الثانية: «الإيمان ينير القلوب»

الوسم الإعرابي: (مب (الإيمان) خب (فع (ينير) مف (القلوب)))

الإيمان = مبتدأ

ينير القلوب = خبر

ينير = فعل

القلوب = مفعول به

المستوى الأول : الإيمان (مبتدأ) + ينير القلوب (خبر)

المستوى الثاني: ينير (فعل) + القلوب (مفعول به)

وجدير بالذكر أن المصطلح بنك الشجرات treebank يستعمل للمدونة الموسومة إعرابياً. وقد يرجع سبب هذه التسمية إلى ما يسمى «التحليل الشجري» الذي يمثل التركيب بشجرة مقلوبة تتشعب فروعها ثنائياً غالباً (مثلا تبدأ بالجملة، ثم المسند والمسند إليه، ثم مكونات كل منهما، وهكذا). ومن يرغب في الاطلاع على مزيد من المعلومات والأمثلة يمكنه الرجوع إلى الدراسة التي تمت في جامعة ليدز البريطانية (Kais Dukes, Eric Atwell and Abdul-Baqee Sharif، ٢٠١٠)، كذلك الدراسة التي يقوم بها فريق من الباحثين في جامعة كولومبيا الأمريكية Habash ، Nizar, Reem Faraj and Ryan Roth، ٢٠١٠).

(٣) الوسم الدلالي: أي المحددات الدلالية: ويقصد بها تحديد المعنى المقصود من كلمة ما بصورة مختصرة، وهو أمر لا بد منه خاصة للمشتركات اللفظية. فكلمة «عين» قد تأتي للدلالة على العين الباصرة (للإنسان عينان)، وعين الماء (ورد القوم العين ليستقوا)، وعين بمعنى الجاسوس (بث القائد العيون ليرصدوا تحركات العدو)، وعين التوكيد في النحو (هذه القضية عينها).

أساليب الوسم المختلفة:

أما طريقة الوسم فيستحسن اتباع ما يعرف بلغة التعليم المعيارية العامة (General Standard Markup Language: GSML). ولأمثلة تطبيقية على اللغة العربية، راجع الدراسات المشار إليها أعلاه عن مشروعات جامعة ليدز البريطانية وجامعة كولومبيا الأمريكية ودرسات ديوكز وزملائه، Dukes Atwell and Sharif، ٢٠١٠، وحبش وزميليه Habash, Faraj and Roth، ٢٠١٠.

ومن الأمثلة على أساليب الوسم المختلفة استخدام القوس المثلث <> قبل وبعد الجزء المطلوب وسمه. مثلاً لبيان أن النص من الشعر الأموي نكتب: <شعر أموي> أبيات من الشعر الأموي </شعر أموي>، حيث تشير المعلومات بين القوسين المثلثين إلى أن النص من الشعر الأموي، كما تبين المعلومات بداية النص ونهايته. ومن الأمثلة الأخرى استخدام الأقواس المربعة، مثلا [S....S] للدلالة على بداية الجملة ونهايتها.

ولا بد من التأكيد على ضرورة بيان حدود العبارات التي تزيد عن كلمة واحدة للتعرف على بداية العبارة ونهايتها. وهذا ينطبق على التعبيرات الاصطلاحية والأمثال والأفعال المتعدية بحروف وما شابهها إذا أردنا للحاسوب أن يعيننا في حصرها وإحصائها آلياً.

ويمكننا وسم الكلمات بطريقة بسيطة، مثل: س-كتب، ف-كتب للتمييز بين الكلمة بوصفها اسماً وبوصفها فعلاً، وذلك عند غياب برنامج للتحليل الصرفي.

ويذكر باوكر وبيرسون (Bowker and Pearson، ٢٠٠٢) أن هناك نظاماً يسمى معيار تشفير المدونة (Corpus Encoding Standard: CES) اتفق عليه الباحثون في أوربا وأمريكا لتعليم النصوص وتحشيتها، تنقسم إلى ثلاث فئات هي:

(أ) التوثيق، ويقصد به الإشارة إلى معلومات ببليوغرافية عن النص ولغته على سبيل المثال.

(ب) بيانات أساسية: وتشمل معلومات عن مكونات النص، مثل الفقرات والعناوين والهوامش ....، إضافة إلى معلومات عن الرسومات graphics .

(جـ) تحشية لغوية: وتشمل المعلومات اللغوية داخل النص، مثل سمات الكلمات الخاصة بقسم الكلمة part of speech والمعلومات الخاصة بالخطاب والوسم النحوي والوسم الدلالي...

جدير بالذكر أن مفهوم التعليم markup وكذلك نظام لغة الـ SGML مطبقان بصورة معدلة على بعض النصوص المتوفرة على الشابكة، باسم HTML، وXML. ويعني HTML: hypertext markup language «لغة تعليم النص التشعبي»، ويعني XML: extended markup language «لغة التعليم الموسعة». (انظر Bowker and Pearson، ٢٠٠٢:٨٠-٨٩).

ثامناً: أمثلة لمدونات للغة العربية

قبل أن أعطي وصفاً لبعض المدونات، أود الإشارة إلى الدراسة القيمة لعبد الرحمن حاج صالح، أستاذ اللسانيات الجزائري الذي يعتبر بحق أول من دعى إلى إنشاء مدونة للغة العربية تسمى ذخيرة اللغة العربية منذ عدة عقود. وقد نشرت الدراسة في مجلة اللسان العربي التابع لمكتب تنسيق التعريب في العدد ٤٨ (١٩٩٩). ونورد فيما يلي أهداف المشروع ومواصفاته كما وردت في الدراسة المذكورة:

«**أ- أهداف المشروع:**

يرمي مشروع الذخيرة اللغوية العربية إلى إنجاز ما يلي:

١. بنك آلي للغة العربية المستعملة بالفعل (بنك نصوص).

٢. معجم آلي جامع للغة العربية مع المقابل الفرنسي والإنكليزي يستخرج من البنك الآلي المذكور (معجم مفردات).

ب - مواصفات المشروع:

سينجز البنك الآلي (أو الحاسوبي) للمعطيات النصية انطلاقاً من الاستعمال الحقيقي للغة العربية ليضم:

١- المؤلفات ذات القيمة الكبيرة في الآداب والعلوم والتكنولوجيا وغيرها، القديمة منها والحديثة.

٢- المحاضرات الجامعية القيمة المنشورة.

٣- المقالات ذات القيمة المنشورة في المجلات الأدبية والعلمية والبحوث القيمة المعروضة في الندوات والمؤتمرات والموائد المستديرة وغيرها.

٤- جميع المعاجم العربية والمزدوجة اللغة القديمة والحديثة (مثل لسان العرب والمعجم الكبير الحديث وغيرهما). والغرض من بنك النصوص الآلي هو أن يكون قاعدة معطيات (أي بيانات) دائمة بحيث تقبل الزيادة والتصحيح على الدوام بسبب تطور المعلومات من خلال الاستعمال الحقيقي للغة العربية وبالتالي أن تصير المصدر الأساسي لإنجاز المعجم الجامع للغة العربية الذي سيحرره العلماء وخاصة أعضاء المجامع العربية وإنجاز العدد الكبير جداً من الدراسات والبحوث في اللغة العربية...» (عبدالرحمن الحاج صالح، ١٩٩٩: ٢). ولكن هذا المشروع الطموح لم يجد طريقه إلى التنفيذ حسب علمنا، لأسباب عديدة يتعلق بعضها بالمشروع وطبيعته[١] وبعضها بعدم توفر الموارد البشرية والمادية اللازمة.

---

(١) المحرر: قد يكون العائق أهداف المشروع ومواصفاته الطموحة التي لا تتواءم مع طبيعة المدونات في ذلك الوقت ولا حتى في الوقت الحالي؛ ولذلك لا بد من وضع أهداف ومواصفات تراعي الممكن لا المأمول والمرغوب فيه مجردين.

(١) **مدونة مشروع معجم الطلاب (السعودي)** الذي أعدته لجنة من اللغويين السعوديين تشمل محمود إسماعيل صالح، وعلي الخبتي، وعبد الرحمن الشمري، وعبدالله الشلال. وكان الهدف منها تأليف معجم للتلميذ في المرحلتين المتوسطة والثانوية.

وكانت مكونات المدونة:

- المقررات الدراسية.
- القرآن الكريم.
- الحديث النبوي الشريف (رياض الصالحين).
- نماذج أدبية عامة.
- صحف ومجلات شبابية.
- برامج إذاعية ومتلفزة شبابية.
- كتابات الطلاب. نصوص المدونة مصنفة (حسب النوع والمصدر والمستوى).

ومجموع كلمات المدونة ثلاثة ملايين كلمة.

(٢) المدونة اللغوية العربية (المدونة العربية):

تعتبر هذه المدونة أكبر مدونة مفتوحة للغة العربية وأكثرها تنوعاً من حيث النصوص والمصادر. نورد أدناه مقتطفات من النبذة الرسمية عن المدونة من موقعها[١]:

- تصميم المدونة العربية:

إن مدونة مدينة الملك عبدالعزيز للعلوم والتقنية ستكون أكبر وأضخم مدونة لغوية للعربية. وهي في مرحلتها الأولى تسعى لجمع سبعمئة مليون كلمة

---

(١) المحرر: هناك أخطاء إملائية وطباعية وصياغية قمت بتعديلها دون الإشارة إليها.

وسوف يزداد حجمها إلى أن تصل إلى بليون كلمة في مرحلة لاحقة إن شاء الله. راعى تصميم المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية عدة معايير خارجية لاختيار نصوص المدونة تعتمد على خمس ركائز أساسية هي: البعد الزمني، والبعد الجغرافي، والوعاء المعلوماتي، و المجال المعرفي، والتصنيف الموضوعي. إضافة إلى هذا فإن المدونة في مرحلتها الحالية هي للنصوص المكتوبة والكاملة فقط ولا تحوي أي نصوص منطوقه مثل الحوارات التلفزيونية أو الخطابات السياسية أو أي نصوص غير مكتملة مثل فصل من كتاب أو جزء من مقال.

- البعد الزمني:

أول المعايير التي تم مراعاتها هو عامل الزمن أو البعد التاريخي والذي يمتد من عصر ما قبل الإسلام وحتى عصرنا هذا والذي أثر على الوعاء أو الصورة التي ظهر فيها النص وكذلك على حجم النصوص المطلوب جمعها لكل فترة زمنية والذي كان على هيئة دالة أسية تتناسب مع تطور المعارف والعلوم والتدوين الخاص باللغة العربية بحيث يزداد عدد كلمات المدونة كلما قربنا من العصر الحديث ويزداد تنوع صورها معه كذلك.

- البعد الجغرافي:

ويقصد به المكان الذي صدر منه النص. ولأن المدونة تعنى باللغة العربية بمجموعها وتحاول أن تكون ممثلة للغة ومتغيراتها فإنه لم يتم تحديد بلد عربي بعينه لجمع النصوص بل إن المطلوب هو تنوع البلدان والكّتاب أو المؤلفين. ويسعى تصميم المدونة لجمع النصوص من جميع البلدان والمناطق العربية قديما وحديثا، كما أن التصميم لا يمنع ضم أي نص مكتوب باللغة العربية من أي بلد كان بحيث لا يطغى بلد أو منطقة على بقية البلدان أو المناطق قدر الإمكان.

- أوعية النشر:

هذا هو البعد الثالث لتصميم المدونة. حيث تم اختيار ما يناسب كل فترة زمنية وما كان سائداً فيها من علوم ومعارف وما كان أكثر انتشاراً وتداولاً بين الناس من أوعية للنشر وما تكون لغته مناسبة ورصينة. فمثلاً تم استبعاد المنتديات الحوارية وصفحات الإنترنت الخاصة والتي يغلب عليها هذا الوقت اللهجات الدارجة ولا تتقيد باللغة العربية الفصحى. وتم اختيار عشرة أوعية للنشر وهي المخطوطات المحققة، الصحف، المجلات، الكتب، الرسائل الجامعية، الدوريات المحكمة، الإصدارات الرسمية، وكالات الأنباء، الإنترنت والمناهج الدراسية. وتم اختيار هذه الأوعية بناء على انتشارها وتأثيرها و رصانة لغتها. وكل فترة زمنية من فترات المدونة لها ما يناسبها من هذه الأوعية.

- المجال المعرفي:

يندرج تحت كل وعاء من الأوعية المختارة مجالات مناسبة له، تحدد مجال النص وسمته العامة. ففي الصحف على سبيل المثال هناك مجالان رئيسان هما الأخبار والمقالات. وفي المخطوطات المحققة وفي الفترة التي كتبت فيها هذه المخطوطات كان هناك مجالات عامة مثل العقائد والفقه وأصوله وعلوم اللغة وغيرها بما يناسب كل فترة. وينطبق هذا على كل وعاء من الأوعية. وهذا البعد يعطي المدونة فرصة أكبر لإيضاح الاختلافات بين كل مجال وآخر وفترة وأخرى كما يوضح أيضاً تنوعها وتمثيلها للغة بشكل أكبر.

- التصنيف الموضوعي:

يندرج تحت كل مجال من المجالات المخصصة للأوعية عدة مواضيع تفصل المجال وتوضح تنوعاته الأدق وتظهر الثراء والتنوع في كل مجال ووعاء. ففي وعاء الصحف وتحت مجال الأخبار هناك عدة مواضيع مثل الأخبار

الاجتماعية، الأخبار السياسية، الأخبار الرياضية، الأخبار الاقتصادية... وفي وعاء المخطوطات المحققة وتحت مجال أصول الفقه هناك عدة مواضيع مثل أصول الفقه الشافعي، أصول الفقه الحنبلي، أصول الفقه المالكي، أصول الفقه الحنفي وأصول الفقه الإثنا عشري. يعطي هذا البعد تنوعاً وثراءً وتخصصاً أكثر لنصوص المدونة مما يجعل الفائدة أكبر منها بحيث يجد الدارس ما يناسبه لاختياره ودراسته ومقارنته.

معلومات عامة عن المدونة:

عدد الكلمات الكلي = ٧٣٩١١٩٠١١

عدد الكلمات بدون تكرار = ٧٤٦٤٣٩٦

العدد الكلي للنصوص = ٩٥٠٤٧٨ نصاً

العدد الكلي للمؤلفين = ١٩٠٠ مؤلفاً» (المصدر: الموقع الرسمي للمدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية www.kacst.org.sa). وقد طورت المدينة مؤخراً برنامجاً لتحليل المدونات اللغوية باسم «الغواص»، وهو متاح على الشابكة (الإنترنت) في الموقع التالي (http//sourceforge.net/projects/kacst-acptool).

(٣) المدونة اللغوية العربية الدولية:

من مشروعات المدونات العربية الطموحة المدونة العربية الدولية International Corpus of Arabic التابع لمكتبة الإسكندرية، وهي مدونة للعربية المعاصرة تطمح إلى بناء مدونة قوامها مئة مليون كلمة. ونورد هنا مقتبسات من التعريف الرسمي للمدونة[١]:

- هدف المدونة اللغوية العربية العالمية:

(١) المحرر: هناك القليل من الأخطاء الإملائية والطباعية والصياغية قمت بتعديلها دون الإشارة إليها.

... بناء مدونة لغوية للعربية المعاصرة تحوي ١٠٠ مليون كلمة محللة صرفياً ونحوياً ودلالياً، وقد روعي فيها أن تكون ممثلة لقطاع إقليمي كبير من الدول الناطقة باللغة العربية المعاصرة وعاكسة بشكل حقيقي وواقعي لأنماط استخدام اللغة العربية المعاصرة في أنحاء العالم العربي. بمجرد الانتهاء من بناء المدونة ستكون أول مدونة محللة ومتاحة كمورد لغوي للباحثين بصفة عامة والباحثين اللغويين بصفة خاصة لتفيد في وصف نظريات اللغة من خلال الاستخدام الواقعي للكلمات.

تخطيط المدونة اللغوية العربية العالمية:

لقد روعيت العديد من الأمور المرتبطة ببناء المدونة مثل التمثيل الجيد للنصوص في العربية المعاصرة والتنوع في فئات النصوص ومحتواها والتوازن بين كل فئة من النصوص وحجم الكلمات المجمعة في كل فئة من فئات التجميع. عند النظر إلى تمثيل العربية المعاصرة داخل المدونة نجد أن الاهتمام الأساسي هو التغطية والتمثيل الواقعي لمختلف المصادر من كل المجتمعات العربية. فشملت المدونة عدداً من المصادر والفئات المختلفة للنصوص وذلك بهدف تحقيق شروط التمثيل الجيد ومدى انتشار المصدر أو الفئة، والتوازن بين كل مصدر وكل فئة، وحجم الكلمات في كل مصدر وفئة.

ونجد أن تصميم المدونة اعتمد بالأساس على البدء بحصر المصادر المختلفة، وداخل كل مصدر تم إدراج الفئات المميزة له. وقد تم حفظ النصوص داخل المدونة بطريقة هرمية من خلال تسمية النصوص بطريقة توضح العديد من المعلومات مثل المصدر والفئة وتاريخ النشر.

لقد تم الأخذ في الاعتبار العديد من الأمور عند تجميع المدونة مثل عدد الفئات المتضمنة داخل المدونة، وعدد النصوص داخل كل فئة من هذه الفئات،

بالإضافة إلى متوسط عدد الكلمات داخل كل نص. تبعاً لطبيعة كل مصدر من مصادر التجميع.

- تصميم المدونة اللغوية العربية العالمية:

- يوجد أربعة مصادر أساسية: الصحافة والمقالات الإلكترونية والكتب والدراسات الأكاديمية. المصدر الخاص بالصحافة منقسم إلى ثلاثة مصادر فرعية: الجرائد والمجلات والصحافة الإلكترونية.

- يوجد إحدى عشرة فئة على مستوى المدونة: العلوم الإستراتيجية والعلوم الاجتماعية والرياضة والدين والأدب والعلوم الإنسانية والعلوم الطبيعية والعلوم التطبيقية والفنون والثقافة والسير الذاتية والنصوص المتنوعة.

- يوجد أربع وعشرون فئة فرعية: سياسة وقانون واقتصاد واجتماع ودين إسلامي ودين مسيحي وأديان أخرى ودين مقارن وقصص وشعر ونثر ودراسات لغوية وأدبية وطب وهندسة وزراعة وتكنولوجيا وعلم الأحياء وعلم الفيزياء وعلم الفضاء وعلم الجيولوجيا والبيئة وعلم الكيمياء وعلم النفس وعلم الفلسفة وتاريخ.

- يوجد أربع فئات فرعية من فئة القصص الفرعية: روايات وقصص قصيرة وقصص أطفال ومسرحيات.

- تغطي المدونة جميع المنشورات داخل الوطن العربي وكذلك بعض المنشورات العربية المنشورة خارج الوطن العربي.

تحليل المدونة اللغوية العربية العالمية:

تشمل هذه المرحلة حالياً التحليل الصرفي لكل كلمة موجودة داخل المدونة، وقد تم في هذه المرحلة تحليل المدونة بطريقة آلية مبنية على بعض الطرق الإحصائية وبعض القواعد اللغوية بالاعتماد على أحد المحللات الصرفية

الشهيرة – تيم باك والتر (Tim Buckwalter) – حيث يوضح التحليل الصرفي عدداً من المعلومات كالسوابق واللواحق وقسم الكلمة وساقها وجذعها وجذرها ووزنها الصرفي بالإضافة إلى نوع الكلمة من حيث الجنس والعدد والتعريف تبعاً للسياقات المختلفة للكلمات داخل كل نص» (المصدر: المدونة العربية الدولية International Corpus of Arabic التابع لمكتبة الإسكندرية).

(٤) مدونات اللغة العربية الأخرى:

إن أفضل حصر للمدونات العربية إلى عهد قريب هو ما قامت به الباحثة لطيفة السليطي من جامعة ليدز (Al-Sulaitie، ٢٠١٠) غير أن القائمة مكتوبة باللغة الإنجليزية. وسنورد في نهاية المبحث أسماء المدونات اللغوية العربية المشهورة وتشمل ترجمة لقائمة لطيفة السليطي، متبوعة بقائمتين إضافيتين نلحقهما بالقائمة المذكورة، وتشمل كذلك عدداً من المدونات الموسومة.

تاسعاً: متطلبات التعامل مع المدونات

تتطلب الإفادة المثلى من المدونات ما يلي:

– محرك بحث search engine: وهذا أبسط برنامج يفيد الباحث في العثور على الكلمات في سياقات مختلفة. وينبغي التنبيه إلى أن البرنامج قد يكون بسيطاً بحيث ينظر إلى الكلمة بمعناها الحاسوبي (حسب الشكل فقط)، وهو البرنامج المتوفر حالياً مع «مدونة اللغة العربية» وبرنامج «غواص» التابعين لمدينة الملك عبدالعزيز للعلوم والتقنية بالرياض، أو يكون متطوراً بحيث يشمل تحليلاً صرفياً، حيث يمكن البحث بالجذر مثلاً أو بالجذع، ويورد البرنامج الكلمة في صورها المختلفة. وهذا النوع الأخير متوفر مع مدونة اللغة العربية التابعة لجامعة بريغهام يونغ (Brigham Young University: BYU) في ولاية يوتا الأمريكية.

- برنامج المكشاف السياقي concordancer، لإعداد الكشاف السياقي
- أي قائمة بألفاظ النص/المدونة في سياقاتها (ما يسمى بالإنجليزية Key Word in Context: KWIC)، بترتيبات مختلفة:

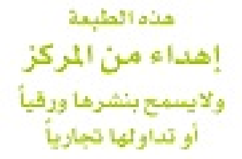
١- الأصل: الكلمة المفتاحية key word مسبوقة ومتبوعة بعدد من الكلمات، حسب ورودها في المدونة. ويمكننا أن نطلب من البرنامج أن:

٢- يجعل ترتيب أسطر الكشاف بناءً على الكلمة المفتاحية وما يسبقها مباشرة <... قرأت كتاباً .....، ... عندي كتاب...> . ويسمى الترتيب الموجه يميناً right sorted.

٣- أو بناء على الكلمة المفتاحية ثم ما يتبعها مباشرة <... كتاب جديد...، كتاب يتحدث عن ...> ويسمى الترتيب الموجه يساراً left sorted. ولا بد من الإشارة إلى أن ربط الكلمتين «يمين» و«يسار» بالسابق واللاحق يختلف في العربية عن المفهوم في اللغات الأوربية التي تتجه الكتابة فيها من اليسار إلى اليمين.

ويمتاز المكشاف السياقي بإمكانات مهمة كثيرة، يجعله يختلف عن محرك البحث التقليدي، مثل إحصاء التكرار وترتيب كلمات المدونة وفق شيوعها (تنازلياً أو تصاعدياً، أي بدءاً بالأشيع أو بالأقل شيوعاً)، إضافة إلى العمل الأساسي له، وهو إيراد الكلمات في سياقاتها.

ومن أمثلة المكشاف السياقي برنامج Aconcorde الذي أعده مجموعة من الباحثين في جامعة ليدز البريطانية، ويعمل مع النصوص العربية والإنجليزية، وبرنامج الغواص alkhawas اللذان صُمما للباحث العربي، وبرنامجا Monoconc، Sketch Engine اللذان يعملان بسلاسة نسبية مع النصوص العربية، غير أن تعليماتهما باللغة الإنجليزية. وبرنامج Word Smith، وهو برنامج متطور من إعداد مطبعة جامعة أكسفورد Oxford University



Press، ولكنه لا يعمل بسلاسة مع النصوص العربية (لمزيدمن التفاصيل عن البرمجيات،انظر المجيول في هذا الكتاب، الجدول ٣٦،الصفحتين ٢٤٤ و٢٤٥).

- برنامج محلل صرفي morphological analyzer للتعرف على الصورة الأساسية للكلمة - الجذع- وكذلك الجذر والوزن)، وكذلك للفصل بين السوابق واللواحق المتصلة بالكلمات، مثل بعض حروف الجر والعطف المتصلة بالكلمة، وغير ذلك. مثلاً كلمة «فهم» قد تكون ف+هُم (الضمير)، أو ف+هَمَّ (الفعل)، أوالفعل فَهِمَ ، أو الاسم فَهمٌ. و الفعل «يكتبون» تعاد إلى جذعها: كَتَبَ ، أو إلى جذرها : ك ت ب والوزن فَعَلَ.

- برنامج لتحديد قسم الكلمة part of speech الذي تنتمي إليه الكلمات، ما يسمى بالإنجليزية POS tagger. وفي غياب ذلك يحتاج الباحث إلى عمل ذلك يدوياً إذا كانت هناك حاجة لذلك. مثال ذلك كلمة «كاتب» قد تكون اسم فاعل أو فعلاً ماضياً. وكلمة «عين» قد تكون اسمًا مفردًا، أو جمعًا (كما في الحور العين)، أو فعلًا (عَيَّنَ فلاناً) فلا بد من وسم كل كلمة لبيان قسم الكلام الذي تنتمي إليه. وهو أمر مطلوب في كثير من البحوث اللغوية.

- برنامج تشكيل آلي يعتمد على الإعراب parsing أي تحديد وظيفة الكلمة النحوية. ويسمى برنامج الإعراب parser (أي المُعرِب).

وللباحثة لطيفة السليطي دراسة مسحية شاملة لأنواع البرمجيات المتاحة للتعامل مع اللغة العربية على مستوى المكشاف والتحليل الصرفي والتشكيل الآلي. (انظر www.leeds.ac.uk/latifa/survey.htm). ومن الأعمال الجديرة بالاطلاع والإفادة الدراسة التي نشرها مروان البواب (٢٠١٢) بعنوان «محركات البحث في النصوص العربية »، ففيها مناقشة مستفيضة للجوانب المهمة في محركات البحث في النصوص العربية ومشكلاتها. وهناك أيضا

دراسة لصوالحة وأتويل يجريان فيها مقارنة بين بعض برامج التحليل الصرفي للغة العربية (Sawalha، M and Atwell، E، ٢٠٠٨).

- الوسم اليدوي manual tagging لإعطاء معلومات عن الكلمات والتراكيب والجمل التي يحتاج إليها الباحث (مثل حدود التعبيرة الاصطلاحية أو العبارات المسكوكة (كما يسميها البعض)، أي وضع علامة تدل على بداية العبارة وأخرى للدلالة على نهاية العبارة. ويلاحظ أن معظم المعلومات التداولية والخطاب تحتاج إلى وسم يدوي، فمعظم هذه المعلومات لا يمكن القيام به آلياً لعدم وجود برمجيات يمكنها القيام بذلك دون تدخل بشري.

- تطوير برنامج للوسم الآلي: هناك برنامج مفتوح المصدر open source يسمى بوابة gate لمساعدة الباحث اللغوي في إعداد برمجيات للوسم الآلي. (الموقع: www.gate.ac.uk). يصفه الموقع بأنه: «مصدر مفتوح لإيجاد حل دورة حياة كاملة لمعالجة النصوص a full-lifecycle open source solution for text processing».

عاشراً: مجالات الإفادة من المدونات اللغوية

أولاً: بعض أنواع التحليل في لسانيات المدونات:

(أ) إيجاد الترابط بين الجوانب اللغوية المختلفة:

١- معجمي-معجمي: أي كلمة مع مصاحبات لفظية، مثل باقة ورد، قطيع ماشية، مواء الهر.

٢- معجمي-تركيبي: أي كلمة تتطلب تركيبًا ما، مثلاً: الفعل الذي ينصب مفعولين أصلهما مبتدأ وخبر، والفعل الذي يتبعه مصدر مؤول أو أن+فعل...

٣- تركيبي-تركيبي: أي تركيب يتطلب تركيباً معيناً، مثل الخبر شبه الجملة الذي يتطلب وقوعه مقدماً على المبتدأ، كما في: في الدار عائلة كبيرة.

(ب) إيجاد ترابط لغوي-غير لغوي: أي العلاقة بين ألفاظ وتراكيب وأساليب لغوية ترتبط بمؤلف، أومنطقة جغرافية أو بلغة ولهجة ما، أو ظواهر لغوية ترتبط بجنس أدبي (شعر أونثر مثلاً)، أو ظواهر لغوية (كالجمل الاسمية أو البناء للمجهول) وارتباطها أو كثرة ورودها في تخصص معين (الطب أو الهندسة أو النقد). ومن أمثلة الترابط اللغوي-غير اللغوي ارتباط بعض الألفاظ بمنطقة جغرافية ما (كلمة «شغل» في المغرب العربي، مقابل «عمل» في المشرق العربي)، وعبارات ورّاق وسوق الورّاقين وارتباطها بلغة التراث العربي.

(جـ) دراسة أنواع النصوص المختلفة: أي دراسة النصوص الأدبية والعلمية والصحفية والقانونية مثلاً وخصائص كل منها.

ويمكننا أن نصنف أسلوب التحليل إلى نوعين:

(١) تحليل كمي quantitative (إحصائي): مثلاً بالنسبة لكلمة ما، عدد مرات ورودها في النص، ما المصاحبات اللفظية لها وما شيوع كل منها.

(٢) تحليل كيفي/نوعي qualitative: ويعني ذلك محاولة تفسير الظاهرة اللغوية، مثلاً: لماذا تشيع كلمة معينة أو تركيب ما في نوع من النصوص؟، مثل شيوع الضمائر في اللغة المحكية مقارنة باللغة المكتوبة أو شيوع المبني للمجهول في النصوص العلمية والتقنية (الإنجليزية) مقارنة بالنصوص الأدبية (انظر Biber et al.، ١٩٩٨).

ثانياً: أهم مجالات الإفادة من لسانيات المدونات:

تعتبر المدونات اللغوية مصدرًا غنيّاً للبحث اللغوي بشتى أنواعه، بل مصدرًا أساساً له في كثير من الأحيان. وكما أشرنا في مكان سابق من هذا المبحث

كان للمدونات المكتوبة والمحفوظة (الشفوية) دورها في استنباط قواعد كثير من اللغات وأنظمتها الصرفية والصوتية وأساليبها البلاغية ومعاجمها أو استقرائها، كما فعل علماء العربية واللسانيون الغربيون الذين درسوا اللغات الأفريقية والأسترالية ولغات الهنود الحمر. ومن خلال المدونات اللغوية (النصوص) توصل اللسانيون إلى معرفة العلاقات بين اللغات، ومن ثم تصنيفها إلى أسر وفروع، كما حدث في علم اللغة التاريخي.

ولعل من أوائل هذا النوع من الدراسات في العصر الحديث الأعمال المعجمية الإحصائية التي استفاد منها صناع المعاجم بأنواعها المختلفة ومعدو مواد تعليم اللغات، خاصة لغير الناطقين بها. وسنورد فيما يلي نماذج للدراسات في مجالات مختلفة اعتمدت أو يمكن أن تعتمد على المدونات اللغوية.

(أ) الدراسات المعجمية:

١- الدراسة المعجمية وصناعة المعاجم lexicography:

كانت الدراسات المعجمية وصناعتها من أوائل التطبيقات العلمية والعملية للمدونات اللغوية (اليدوية في حينها)، حتى من قبل التدوين الإلكتروني أو الرقمي، كما هو واضح من أعمال مايكل وست Michael West ١٩٥٣ وثورندايك ولورج Thonrndike and Lorge ١٩٧٢.

ونجد شرحاً جيداً لدور المدونات المحوسبة في كتابات سنكلير Sinclair الباحث الرئيس في مشروع مدونة COBUILD الذي تم بالتعاون بين جامعة بيرمنغهام Birmingham وناشر المعاجم كولينز Collins (انظر Sinclair، J.، ١٩٩١).

ويعدّ هذا التوجه من أشيع نماذج الإفادة من المدونات اللغوية المطبوعة والمحوسبة. فقد كانت جميع قوائم شيوع الألفاظ في اللغات المختلفة مبنية على مدونات مختلفة. من هذه القوائم الرائدة في الإنجليزية قوائم مايكل وست

Michael West وثورندايك ولورج Thorndike and Lorge. وفي العربية قوائم بريل ولانداو وفاخر عاقل وداود عبده (انظر داود عبده، ١٣٩٩هـ)، فقد قام هؤلاء بالإحصاء اليدوي للكلمات في نصوص مطبوعة للتعرف على الكلمات الشائعة فيها. كذلك قام بعضهم بأبعد من ذلك، حيث قام مايكل وست بإحصاء تكرار المعاني المختلفة للألفاظ semantic count (انظر Michael West، ١٩٥٣).

تحديد مفهوم الكلمة والوحدة المعجمية والمصطلحات ذات العلاقة:

من القضايا المهمة التي يجب التنبه لها عند الحديث عن الدراسات المعجمية التفريق بين المصطلحات ذات العلاقة بالكلمات وتعريفاتها المخلتفة، ومنها ما يلي:

١- الكلمة الحاسوبية: مجموعة من الحروف (أو حتى رمز واحد) مسبوقة بفراغ ومتبوعة بفراغ، مثلاً: الأرقام والفاصلة أو علامة الاستفهام إن كانت مسبوقة ومتبوعة بفراغات، والكلمات «ويكتبون»، «فكتبناهم» على الرغم من اشتمالها على عدة وحدات صرفية.

٢- الكلمة الفعلية/النصية token: وهي الكلمة كما ترد في النص بصور مختلفة (مثلاً الفعل بتصريفاته المختلفة – كتب، كتبنا، كتبا، نكتب، يكتبون... أو الاسم في صيغة المفرد والمثنى والجمع – تلميذ، تلميذان، تلاميذ...). ويستعمل المصطلح أحياناً ليعني الكلمة الحاسوبية إن كانت لغوية.

٣- الوحدة المعجمية lexeme: قد تكون كلمة واحدة أو عبارة تتكون من أكثر من كلمة، كما في التعبيرات الاصطلاحية (مثلاً: ضرب أخماساً في أسداس، بمعنى احتار).

٤- الجذع stem، lemma: وهي الكلمة المجردة من الزوائد الصرفية كعلامات الجمع أو التثنية أو الدلالة على الزمن في الأفعال. وهو للاسم الكلمة في صيغة المفرد (وجمع التكسير، أحياناً)، مثل كاتب، مكتوب، وللفعل صيغته في الماضي للمفرد المذكر، مثل: كتب، قال. وهو ما نجده في مداخل المعاجم المرتبة حسب الكلمات (أو النطق).

٥- الكلمة النوعية type: وهي الكلمة النوعية أو الكلمة المختلفة: هي الكلمة في أية صورة بغض النظر عن تكرارها.

٦- الجذر root: (في العربية، مثل: ك ت ب ، ق و ل) والمكون الأساس للكلمة في اللغات الأوربية (من أمثلة ذلك الجذور اليونانية واللاتينية، مثل bio، geo، logy)، إضافة إلى الصورة الأبسط للكلمة.

٧- تجريد الكلمة من الزوائد (تجذيع lemmatization): ويعني ذلك في العربية تحويل الكلمة إلى صورتها الأساس (الجذع) مع حذف الزوائد السابقة، مثل حرف الجر «ب» والعطف «ف» واللاحقة مثل الضمائر المتصلة «ه» و«ها». فالكلمة الحاسوبية «فهم» تجرد إلى «فهم» الاسم والفعل، وكذلك إلى «ف» +«هم». والكلمة الحاسوبية «بالقلم» نحذف منها الباء لتصبح «القلم».

٨- قائمة الكلمات المستثناة stop list: عند طلب الكشاف السياقي أو قوائم الشيوع في المدونة، هناك كلمات لانحتاج أحياناً إلى معرفة شيوعها (وهي ماتسمى بالكلمات الوظيفية، مثل الضمائر وحروف الجر أو حروف العطف)، لكثرتها ولأننا نعرف سلفا أنها أكثر الكلمات شيوعاً في اللغة. في هذه الحالة يمكننا أن نعطي للبرنامج الإحصائي قائمة بهذه الألفاظ ليتجاهلها الحاسوب. وهذه هي التي تسمى قائمة الكلمات المستثناة stop list .

٩- شيوع الألفاظ word frequency: للتعرف على مدى شيوع كلمة معينة أو تكرارها نقوم بعملية إحصاء الكلمات word count، مع ملاحظة التعريفات المختلفة لمفهوم الكلمة المذكورة أعلاه. والأصل في ذلك النظر إلى تكرار الكلمة النوعية أو الجذع بغض النظر عن صور ورودها فعلياً (الكلمة الفعلية). ويجب التنبيه إلى عدم الخلط بين أنواع الكلمة بمفاهيمها المختلفة المذكورة، وكذلك بين الكلمة والجذر. فالجذر «غ ف ر» تشتق منه جذوع أو كلمات مختلفة تختلف من حيث شيوعها ومعانيها (غفر، استغفر، غافر، غفّار، غفور، مغفرة).

١٠- شيوع المعاني semantic count: يشير هذا المصطلح إلى إحصاء المعاني المختلفة للكلمة متعددة المعاني polysemous، وينطبق ذلك خاصة على المشتركات اللفظية homonyms، مثل «عين» الباصرة ومنبع الماء وأداة التوكيد، فلكلّ من هذه المعاني تكرارها الخاص بها. ونجد مثالاً لإحصاء شيوع المعاني المختلفة للألفاظ في كتاب مايكل وست الريادي الذي أشرنا إليه سابقاً (Michael West، ١٩٥٣).

مجالات البحث التي تتعلق بالجانب المعجمي للغة:

من أهم هذه المجالات مايلي:

١- ما مدى شيوع كلمة ما، بصورة عامة أو في نصوص مختلفة النوع مثلاً؟

٢- ما نسبة شيوع كلمة ما في نصين مختلفين؟

عند إجراء هذا النوع من المقارنة علينا أن نأخذ بعين الاعتبار حجم كل نص. لذلك يفرق الخبراء بين الإحصاء أو العد العام/الخام raw، أي ذكر التكرار في كل نص بغض النظر عن عدد كلماته. فالمفروض أن نلجأ إلى حساب التكرار النسبي normed count، أي نذكر تكرار الكلمة في

كل ألف كلمة مثلا من النصين أ و ب، كأن نقول إن كلمة «مسألة» ترد ٥٠ مرة في كل ألف كلمة من النص أو ترد ٥٧ مرة في كل ألف كلمة من النص ب.

٣- ما مدى شيوع معاني كلمة ما (الإحصاء الدلالي semantic count) في نص ما أو في نصوص مختلفة؟ نجد مثلاً أنه في الاستعمال العام ترد كلمة «عين» بمعنى البصر أكثر منها بمعنى عين الماء. كذلك نجد أن كلمة «باب» بمعنى المدخل قد يكون أكثر شيوعاً منها بمعنى فصل من كتاب.

٤- هل للكلمة ارتباط بكلمات معينة (المصاحبات اللفظية - سابقة أو لاحقة)؟ مثلاً: «جماعة من أهل العلم» و«فريق من اللاعبين» و«قطيع من الماشية».

٥- هل للكلمة ارتباط بلهجة أو نوع من الاستعمال اللغوي genre أو غير ذلك من العوامل غير اللغوية؟ (مثلاً: كلمة تكوين بمعنى التدريب، والشغل بمعنى العمل في المغرب العربي).

٦- ما مدى شيوع المصاحبات اللفظية المعينة، بصورة عامة أو في لهجات أو أنواع من الاستعمال اللغوي؟

٧- ما مجالات استعمال المترادفات (وسيم، جميل أو كبير، ضخم) وتوزيعاتها، حسب نوع النص مثلاً؟

٨- الاختلاف بين استعمالات الكلمة في حالتي الإفراد والجمع (مثلا كلمة «حبل» و«حبال» في القرآن الكريم).

٩- التعرف على التعبيرات الاصطلاحية (المسكوكات) ومدى شيوعها عامة أو في أنواع مختلفة من النصوص. ويشمل ذلك الأفعال المتعدية بحروف. (انظر مثلاً وفاء كامل، ٢٠٠٧؛ ومحمد الحناش، ١٩٩٦).

١٠- دراسة الجذور والحروف العربية من حيث تكرارها وشيوعها. ويعتبر علي حلمي موسى في دراسته (بالتعاون مع إبراهيم أنيس وعبدالصبور شاهين) والخاصتين بمعجمي لسان العرب والصحاح رائداً في هذا النوع من الدراسات (انظر قائمة المراجع) . ومن الأعمال الرائدة في العالم العربي أيضاً (المعجم الحاسوبي: إحصاء الأفعال العربية في المعجم العربي) من إعداد محمد المراياتي ويحي ميرعلم ومحمد حسن طيان. ولكن يلاحظ أن هذه الأعمال عاملت بعض المعاجم العربية بوصفها مدونات في إنجاز الدراسات المشار إليها (انظر الدراسة القيمة لعبدالرحمن حاج صالح في قائمة المراجع، حيث يذكر عدداً من فوائد المدونة اللغوية الإضافية. ومن الكتب التي عالجت بإسهاب كثيراً من القضايا المعجمية كتاب سنكلير الذي كان من المشرفين على مشروع COBUILD ومحرراً لمعجم BBC للغة الإنجليزية).

صناعة المعاجم:

يلاحظ أن إعداد المعاجم (أحادية اللغة وثنائيتها) يستخدم المدونات لأغراض مختلفة، مثل اختيار ألفاظ المعجم (المداخل) وكذلك حصر معاني الألفاظ (بناءً على الكشاف السياقي والنصوص الأصلية- الشواهد) وفي التمثيل لاستعمالات المداخل المختلفة (من واقع المدونة). من المشروعات الرائدة معجم Collins COBUILD English Language Dictionary. ومن المشروعات الحديثة معجم كامبريدج للمحتوى الأكاديمي Cambridge Academic Content Dictionary الذي صدر عام ٢٠٠٩ والذي اعتمد مصادر رسمية وشبه رسمية للمعايير والاختبارات المقننة وماشابه ذلك في اختيار مداخله. كذلك يستخدم الحاسوب أحيانا في التحكم في لغة التعريف، كما فعلت

لونغمان في معجمها Longman Dictionary of Contemporary English، حيث حددت هذه اللغة بالألفي كلمة الأكثر شيوعاً في اللغة الإنجليزية.

نجد الشيء نفسه في معجم كامبريدج للمحتوى الأكاديمي المذكور آنفا والموجه للطلاب كما يبدو، حيث يذكر محرره أن «تعريفات المعجم كتبت باستخدام مفردات كامبريدج للتعريفات Cambridge Defining Vocabulary، وهي قائمة من ٢٥٠٠ كلمة شائعة والتي يعرفها الطلاب. وقد تم تطوير هذه القائمة بالإفادة من الأجزاء المتعلقة بالإنجليزية الأمريكية من مدونة كامبريدج الدولية Cambridge International Corpus، وهي قاعدة بيانات تحوي أكثر من مليار كلمة مكتوبة ومنطوقة» (مقدمة المعجم المذكور: vii).

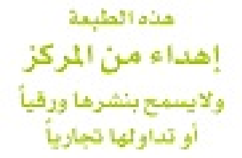
٢- دراسات المصطلحية (علم المصطلح terminology):

أي التعرف على المصطلحات في المدونة، ومعرفة شيوعها في النصوص المختلفة. (انظر Bowker، Lynne and Pearson، Jennifer، ٢٠٠٢ الفصلين ٨ و٩)، وانظر كذلك الدراسة التي قدمها عبدالمحسن الثبيتي في: الندوة الدولية الأولى عن الحاسب واللغة العربية: الأوراق البحثية، ٢٠٠٧؛ والدراسة المقدمة من عز الدين غازي في الندوة ذاتها. ولعل أفضل كتاب شامل بالعربية يعالج دور المدونات اللغوية في دراسة المصطلحات الكتاب الذي ترجمته ريما بركة: علم المصطح: مبادئ وتقنيات (انظر ماري-كلود لوم، ٢٠١٢ الفصول الرابع والخامس والسادس).

ب- اللسانية العامة:

أولاً: الدراسات الصوتية:

لعل أفضل مثال لهذا النوع من الأعمال المبنية على المدونات اللغوية ما قام به رائد هذا النوع من الدراسات في هذا المجال علي حلمي موسى، حيث يتحدث عن العلاقة بين الصوامت والصوائت في العربية من الوجهة الإحصائية. يقول



في معرض كلامه: «والدراسة التي نعرضها اليوم تمت على عينتين من القرآن الكريم إحداهما مكية وهي سورة الأعراف وبعض قصار السور والأخرى مدنية وهي سورة البقرة. ونظراً لأن الدراسة تتم على القرآن كما نسمعه وليس كما نكتبه فقد اتبعت الطريقة الصوتية ...، حيث نتبع طريقة قراءة حفص مع الوقوف على رؤوس الآيات» (علي حلمي موسى، ٢٠٠١).

ومن المشروعات المهمة في مجال الأصوات المشروع الذي قام به منصور الغامدي في مدينة الملك عبد العزيز للعلوم والتقنية في الرياض (انظر الغامدي في قائمة المراجع).

ويجدر بنا أن نشير إلى أن أية دراسة إحصائية عربية للجذور في المعاجم هي في واقع الأمر دراسة لأصوات العربية، خاصة الصوامت، لأن الجذر يتكون من حروف (أي أصوات). عليه فإن جميع أعمال علي حلمي موسى المتعلقة بالمعاجم العربية المعروفة تعتبر دراسات صوتية، إضافة إلى كونها معجمية.

ثانياً: الدراسات النحوية والصرفية:

يمكننا أن ندرس هذا الموضوع من زوايا عديدة، مثل شيوع التراكيب بعامة أو تراكيب معينة في أنواع من النصوص المختلفة. (انظر لأمثلة على ذلك: الفصلين الثالث والرابع من Biber, Conrad and Reppen، ١٩٩٨، وكذلك الفصول ١٠، ١١، ١٢ ، ١٦ في Aijmer and Altenberg eds، ١٩٩١).

أمثلة للبحوث في المجالات النحوية والصرفية:

يبدو أن معظم الدراسات الأولى التي استفادت من المدونات اللغوية كانت الدراسات المعجمية. ولكن هناك توجهاً متزايدًا في استخدامها في الدراسات النحوية والصرفية، خاصة في بعض اللغات مثل الإنجليزية، كما نلاحظ في بايبر وزملائه Biber et al. ١٩٩٨، وماير Meyer، ٢٠٠٢، وأيجمر وألتنبيرغ

Aijmer and Alternberg، ١٩٩١. ونورد أدناه أمثلة لهذا النوع من الدراسات التي يمكن إجراؤها على اللغة العربية:

١- شيوع وتوزيع التراكيب النحوية المختلفة (الجمل الاسمية والفعلية، التركيب الوصفي والإضافي، المبني للمعلوم والمبني للمجهول، والمبني للمجهول الذي يشمل الفاعل في صيغة «بواسطة» أو «مِن قِبَل» فلان أو كذا...) في اللغة بصورة عامة أو في النصوص التي تنتمي إلى مجال معين أو منطقة جغرافية أو فترة زمنية.

٢- ارتباط التراكيب بألفاظ أو تراكيب أخرى معينة (مثل ما نجد في العلاقة بين كان وأخواتها وأسمائها وأخبارها).

٣- ارتباط بعض التراكيب مثل أن المصدرية مقابل المصدر بأفعال معينة، أو ارتباط بعض الأفعال المتعدية إلى مفعولين بتراكيب معينة (مثلاً أصلهما مبتدأ وخبر أو أصلهما ليس مبتدأ وخبرًا).

٤- العوامل التي تؤدي إلى اختيارنا لتراكيب وتفضيلنا لها على تراكيب مشابهة في المعنى أو الوظيفة (كالجملة المبنية للمعلوم والجملة المبنية للمجهول في الأساليب الحديثة خاصة). من أمثلة ذلك كثرة ورود الجمل المبنية للمجهول في النصوص العلمية والتقنية، حيث التأكيد على العمليات والإجراءات بدلاً من الاهتمام بالفاعل. كذلك نجد كثرة الأفعال والصفات في النصوص القصصية، لما لهذه الجوانب من أهمية في سرد الأحداث ووصف الشخصيات في القصص.

٥- دراسة المشتقات والصيغ الصرفية المختلفة ومدى شيوعها في نصوص أو أنواع من النصوص المختلفة، وكذلك للتعرف على الأوزان والصيغ المختلفة، مثل اسم الفاعل وصيغة المبالغة «فعّال» أو المصدر الصناعي

ومدى إنتاجية productivity بعضها (كثرة أو قلة استخدام وزن ما في توليد الكلمات الجديدة مثلاً).

٦- دراسة السوابق واللواحق المختلفة (في الإنجليزية مثلاً: in-، un-، -tion، -ity، -ic، -ism) وشيوعها في النصوص المختلفة. مثلاً أيّ هذه اللواحق أكثر وروداً في النصوص العلمية؟ وأيها أكثر استعمالاً في النصوص الصحفية؟ وفي العربية يمكننا مقارنة شيوع لواحق جمع المذكر السالم والمؤنث السالم مثلاً في جمع الكلمات المقترضة (مثلاً نلاحظ استعمال الألف والتاء في معظم الكلمات المقترضة، مثل كومبيوترات وتلفزيونات وموبايلات).

٧- دراسة شيوع أنواع الأفعال المختلفة، مثل الأفعال الدالة على الحركة (مشى، جرى) والدالة على التغيُّر أو الثبات من جهة، والصيغ الزمنية المختلفة في أنواع النصوص المختلفة. مثلاً في البحوث والرسائل العلمية نجد أن الباحث يستخدم صيغة المستقبل في بعض الفصول والماضي في فصول أخرى.

٨- معرفة نسبة ورود أقسام الكلمة parts of speech المختلفة في المدونة أو في أنواع النصوص المختلفة (مثلاً نسبة الأفعال إلى المصادر أو الأسماء في نص أو نوع معين من النصوص).

٩- لعل من المجالات المهمة في النحو العربي دراسة قضية أقسام الكلام وتصنيف الألفاظ العربية من منظور جديد (مثلاً سبعة أقسام اقترحها تمام حسان، بدلاً من الثلاثة المعروفة (انظر محمود الصيني، ١٤١٢هـ).

١٠- من الدراسات التي لم تلق العناية الكافية، وهي جديرة بالبحث في ضوء لسانيات المدونات الصيغ المختلفة التي تعبر بها العربية عن الزمن

والوجهة tense and aspect، مثل «كان قد فعل، قد فعل، كان يفعل، سوف يكون قد فعل...» وماشابه ذلك (انظر Sieny, Mahmoud، ١٩٨٦).

ثالثًا: التطور التاريخي للغة من حيث المعجم والتراكيب النحوية:

تفيدنا المدونات التايخية أو اللغوية عبر العصور في هذا النوع من الدراسات (انظر Biber, Conrad and Reppen، ١٩٩٨:٢٠٣–٢٢٩). من أمثلة هذه المدونات المدونة التاريخية المشهورة للغة الإنجليزية للفترة مابين ١٦٥٠ إلى ١٩٩٠ ARCHER إعداد باحثين في جامعة أريزونا الشمالية Northern Arizona بالولايات المتحدة الأمريكية، ومدونة هلسنكي Helsinki للفترة من ٨٥٠ إلى ١٧١٠ من تاريخ اللغة الإنجليزية.

إن أي معجم تاريخي للغة العربية لا يستغني عن مدونة تاريخية شاملة وممثلة للعصور المختلفة، تقدم قدراً متوازناً من النصوص المختلفة لكل فترة زمنية (بحسب القرون أو العصور الأدبية)، يحددها المتخصصون في اللغة والأدب العربي، مثل العصر الجاهلي، والعصر الإسلامي، والعصر الأموي....

رابعًا : تحليل الخطاب discourse analysis وتحليل النصوص text analysis:

من المعروف أن هذين الفرعين من اللسانيات يتعاملان مع العلاقات فوق مستوى الجملة، بما في ذلك قضايا الترابط النصي cohesion والدلالي coherence. ومن أمثلة هذا النوع من الدراسات شيوع استخدام الضمائر في النصوص الشفوية مقابل النصوص المكتوبة والمسافة بين الضمير العائد والاسم الذي يعود إليه في هذه النصوص، وكذلك اختلاف الصيغة الزمنية للأفعال في أجزاء البحوث العلمية المختلفة (المقدمة ، المنهج ، النتائج مثلاً) (انظرBiber, Conrad and Reppen، ١٩٩٨:١٠٦–١٣٢؛ و Marcus Callies،

٢٠٠٨؛ و Tottie; ١٩٩١) انظر أيضاً محمود إسماعيل صالح، ٢٠١٤ لأمثلة عربية في هذا المجال.

خامسًا : التداولية pragmatics و أعمال الكلام speech acts:

من الصعب أن نتصور أية دراسة لسانية تعالج الاستعمال اللغوي والعلاقة بين اللغة والوظائف والمواقف الاجتماعية في غياب مدونة مناسبة تشتمل على أمثلة واقعية لهذا الاستعمال اللغوي. ونلاحظ أنه في غياب مدونات لنصوص واقعية قد يلجأ الباحثون إلى ما يسمى بالنصوص المستنطقة elicited (مقابل التلقائية spontaneous). ومن الأمثلة على هذا النوع من الدراسات البحث الذي قامت به عبير الطويل وزملاؤها في الأردن عن «المراوغة /التحاشي hedging» في الخطاب السياسي في العربية والإنجليزية (Taweel، Abeer et al.، ٢٠١١). ومن النماذج من الدراسات في هذا المجال باللغة العربية مجموعة الدراسات المتعلقة بالخطاب والتداولية الواردة في كتاب: دراسات في علم اللغة النصي: مقاربة تطبيقية على مدونة الجزيرة، ٢٠١٢. ولمزيد من المعلومات حول المدونات اللغوية واللسانيات بصورة عامة، نقترح الرجوع إلى الدراسة الحديثة في هذا المجال لصالح العصيمي، ٢٠١٣.

جـ- اكتساب اللغة وتعلمها وتعليمها:

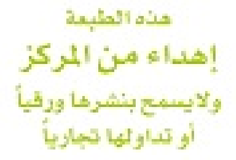
١- اكتساب اللغة الأولى والتطور اللغوي للأطفال:

كانت دراسات النمو اللغوي للأطفال تعتمد على الملاحظات الشخصية لعدد محدود من الأطفال على مدى بضع سنوات في بعض الأحيان؛ فقد كانت محدودة من حيث العينات التي تخضع للبحث. أما مع توفر مدونات لغوية لآلاف الأطفال من أعمار مختلفة فقد أصبح بالإمكان دراسة الموضوع على نطاق أوسع، إضافة إلى إمكانات المقارنة بين مئات الأطفال من عمر واحد أو أعمار مختلفة، مما يعطي للاستنتاجات مصداقية أعلى. كما يفيد ذلك

إذا ما أجريت الدراسة على مدونات للغات مختلفة أن نتعرف على ما يسمى بالظواهر اللغوية العالمية أو الكليات اللغوية language universals بصورة أدق وأفضل. لعل من أهم المدونات اللغوية لمثل هذه الدراسة باللغة الإنجليزية مدونة «CHILDES» من إعداد قسم علم النفس في جامعة كارنيجي ميلون Carnegie Mellon; في الولايات المتحدة الأمريكية (انظر Biber et al.، ١٩٩٨:١٧٢-٢٠٢) حيث يورد المؤلفون أمثلة على الدراسات التي أجريت على التطور اللغوي لدى الأطفال، ويذكرون أمثلة على الملامح المميزة لكتابات التلاميذ مقارنة بالبالغين.

٢- تعليم اللغات:

من أبرز نماذج الإفادة من المدونات في تعليم اللغات المساهمة في اختيار محتوى المادة التعليمية (من مفردات وعبارات وتراكيب) وفي إعداد المراجع المعجمية والنحوية، كما فعل الناشر كولنز Collins في مجموعة المراجع المبنية على مدونة COBUILD تحت هذا المسمى. يضاف إلى ذلك الاستخدام المباشر للمدونات في تعليم مفردات اللغة وتراكيبها. بل إن هناك اتجاهاً حديثاً في تعليم اللغات يسمى المنهج المعجمي lexical approach في تعليم اللغات يعتمد بدرجة كبيرة على المدونات اللغوية (انظر Richards and Rogers، ٢٠١٤). ونود الإشارة إلى أن هناك كتابين صدرا مؤخراً بعنوان: من المدونة إلى صف الدراسة (O'Keefe et al.، ٢٠٠٧)، وأنماط نصية: الكلمات المفتاحية وتحليل المدونات فى تعليم اللغة (M. Scott and Ch. Tribble، ٢٠٠٦)، كذلك خصصت سوزان هنستون (S Hunston، ٢٠٠٢ الفصول ٦ و٧ و٨ من كتابها عن المدونات اللغوية واللسانيات التطبيقية.



د- الترجمة والتحليل التقابلي وتحليل الأخطاء:

١- دراسات الترجمة:

يمكن الإفادة من المدونات المتوازية خاصة في دراسة خصائص لغة الترجمة والمترجمين، وكذلك في دراسات المتقابلات في هذه المدونات. كما أن هناك توجهاً للاستفادة من هذه المدونات في تطوير برامج الترجمة الآلية (كما فعلت شركة IBM المعروفة)، حيث تدرس الاحتمالات الإحصائية للترجمات المتقابلة في هذه المدونات لتطبيقها في الترجمة الآلية. من أمثلة المدونات التي تتعلق بالعربية المدونة التي طورتها جامعة الكويت لهذا الغرض E-A Parallel Corpus «مدونة متوازية إنجليزية عربية» (انظر Al-Ajmi، H.، ٢٠٠٤)، وكذلك المدونة المتوازية لنصوص الأمم المتحدة: إنجليزي-عربي (انظر Hunston، ٢٠٠٢:١٢٣-١٢٧؛ و Olohan، ٢٠٠٤).

٢- التحليل التقابلي contrastive analysis:

من فوائد المدونات المقارنة أو المتناظرة comparable corpora إجراء الدراسات التقابلية بين اللغات وبين التنوعات اللغوية (انظر على سبيل المثال Gilquin,Papp and Diez-Bedman، ٢٠٠٨)، ويندرج تحته الدراسات المقارنة بين اللهجات المختلفة، إضافة إلى المقارنة بين اللغات، وبين لغة المتعلمين ولغة أهل اللغة الأصليين.

٣- تحليل الأخطاء:

قد يكون هذا المجال أول مجال في علم اللغة التطبيقي يستخدم النصوص التي ينتجها الدارسون وهو ما يسمى بمدونة المتعلمين learner corpus. وهناك أعداد متزايدة من مثل هذه المدونات لدارسي اللغات المختلفة من غير أهلها

من خلفيات ومستويات وبيئات مختلفة، مما يتيح للباحث التعرف على الأخطاء الشائعة لدى الناطقين بلغة ما أو بلغات مختلفة، كذلك استنتاج التطور اللغوي لدى المتعلمين. وتحليل الأخطاء بطبيعته يعتمد على أمثال هذه المدونات أساساً، كما هو واضح من أي دراسة في تحليل الأخطاء في شتى اللغات (انظر المبحث الثاني من هذا الكتاب للتفاصيل حول هذا الموضوع).

هـ- مجالات لسانية أخرى:

١- الأسلوبية:

يذكر بايبر وزميلاه أن من أمثلة الدراسات الأسلوبية الدراسات التي تجرى لتحديد هوية المؤلف (المجهولة أو المشكوك فيها) بمقارنة النص بنصوص أخرى معروفة المؤلف، وكذلك الدراسات التي تهدف إلى التعرف على الخصائص الأسلوبية لبعض النصوص (Biber et al.، ١٩٩٨:٢٢٣)؛ وانظر أيضا Crystal، ١٩٩١. كما أن هناك فوائد المدونات في خدمة الكتّاب لتحسين أسلوبهم في التعبير؛ إذ تذكر هنستون أن بعض الباحثين «استخدموا المعلومات المستقاة من مدونة عامة في تقديم المشورة لكتاب الوثائق الرسمية الموجهة للقارىء العادي من حيث الاستعمال الأنسب للغة» (Hunston، ٢٠٠٢:١٣٥). وجدير بالذكر أن هناك دورية تعنى بهذا النوع من الدراسات تسمى الحوسبة الأدبية واللغوية Literary and Linguistic Computing، وقد بلغت سنتها السابعة والعشرين في العام ٢٠١٢.

٢- الأيديولوجية (العقائد الفكرية) والثقافة وعلاقتهما باللغة:

تناقش الباحثة سوزان هنستون Susan Hunston في القسم المعنون «دراسة الأيديولوجيا والثقافة Studying ideology and culture:١٠٩-١٢٣» من كتابها أمثلة لهذا النوع من الدراسات، حيث تقول: إن «المدرسة السائدة

في بحث العلاقة بين اللغة والعقائد (الأيديولوجيا) هي اللسانيات الناقدة critical linguistics أو تحليل الخطاب الناقد critical discourse analysis الذي يدرس اللغة ليس بوصفها منظومة مستقلة ولكن بوصفها شيئا يتدخل intervenes في المجتمع، غالبا بترسيخ فرضيات وقيم ذلك المجتمع ...». وتذكر أن أحد رواد هذا النوع من الدراسات (فاولر Fowler) ذكر أن «هناك ثلاثة جوانب للسانيات الناقدة هي:

- دراسة النصوص في سياق الظروف الاجتماعية التي أنتجت فيها.

- الكشف عن الأيديولوجية المتضمنة implicitly coded وراء الكلام والأفكار المذكورة صراحة overt propositions.

- تحدي الحس العام common sense بالتنبيه إلى أنه كان بالإمكان أن نمثل شيئاً ما بطريقة مختلفة لها مضمون مختلف» (Hunston، ٢٠٠٢:١٠٩). (انظر كتاب سوزان هنستون Hunston, Susan، ٢٠٠٢ حيث تناقش الباحثة المجالات المختلفة من اللسانيات التطبيقية التي تستفيد من المدونات اللغوية. وكذلك انظر مقال «المدونات اللغوية واللسانيات التطبيقية في McEnery et al.، ٢٠٠٦:٨٠-١٢٤).

ولعل من أهم الدراسات التي كتبت في هذا المجال عن اللغة العربية دراسة رجب جمعان الزهراني (٢٠١٣: Alzahrani، Rajab Jamaan :٢٠١٣)، والتي يقول في ملخصها: « تدرس هذه الأطروحة التمظهرات الأيديوجية ideological representations للخطاب السلفي في السعودية من الفترة ١٩٨٠-٢٠٠٠ وتحاول أن تجيب عن االسؤال التالي: ما مدى وشكل تجانس الخطاب السلفي في السعودية في الفترة بين ١٩٨٠ و٢٠٠٠.».

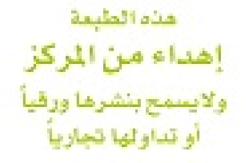
٣- اللسانيات الجنائية forensic linguistics:

من المجالات الطريفة في الإفادة من المدونات «اللسانيات الجنائية forensic linguistics» أي دراسة اللغة لأغراض قانونية وقضائية، مثل قضايا

التزوير وتحديد هوية المجرمين. وتذكر هنستون أن هناك حزمة برمجية تسمى CopyCatch (اضبط النسخ) تقارن نص «المشتبه بهم» بنص من إنتاج غيرهم (النص الضابط control text)، وتذكر أن فاينالي (Finlay، ٢٠٠٢) «تذكر عدداً من الملامح المهمة في تمييز النصوص المشبوهة من النصوص الضابطة» (Hunston، ٢٠٠٢:١٣٣). وجدير بالذكر في هذا السياق ما أشار إليه كارل جيمس في كتابه عن أخطاء المتعلمين (James، ١٩٩٨) من أن تحليل الأخطاء اللغوية يمكن استعماله للغرض نفسه.

و- دراسات ومشروعات اللسانيات الحاسوبية أو تطبيقاتها:

يلاحظ أن عدداً من المدونات المختلفة كان وراءها خدمة البحث اللساني الحاسوبي، كما هو واضح من الدراسات ذات الطبيعة الإحصائية ونظرية الاحتمالات، كما نجده في أعمال الحاسوبيين الذين يعملون على التعرف على الكتابة العربية، وكما هو الحال مع مدونتي CALLFRIEND، CALLHOME للهجة المصرية اللذين كان الغرض منهما تطوير برنامج للتعرف على الكلام، ومدونة الحياة (Al-Hayat Corpus) لأغراض الهندسة اللغوية واسترجاع المعلومات (انظر Al-Sulaiti، ٢٠١٠)، كذلك الأمر مع عدد من مشروعات الترجمة الآلية المبنية على الاحتمالات الإحصائية المعروفة في العالم. ويستطيع الباحث أن يلاحظ ذلك في كثير من دراسات اللسانيات الحاسوبية (للأمثلة العربية في هذا المجال، ينظر نبيل علي وعبد الغني أبو العزم والعناتي وجبر (٢٠٠٧) للدراسات المنشورة والتي قدمت في ندوات تعريب الحاسوب المختلفة). ويذكر الباحثان كيلغاريف وغرينفينستيت أن بوادر الارتباط الرسمي بين المدونات واللسانيات الحاسوبية يرجع رسمياً إلى العام ١٩٩٣ (Kilgarriff, Adam and Grefenstette، ٢٠٠٣).



وقبل أن نختم هذا الفصل، أود التنبيه إلى أن أشمل مصدر للأمثلة المختلفة لدراسات لسانيات المدونات هو المجلة الدولية المعروفة باسم: المجلة الدولية للسانيات المدونات International Journal of Corpus Linguistics.

الخاتمة

كما أشرنا من قبل في بداية هذا المبحث فإن المدونات اللغوية ولسانيات المدونات اللغوية حقل جديد نسبياً في مجال العمل اللساني. وقد حاولنا في هذا الفصل أن نلقي الضوء على أهم جوانب هذا الموضوع، بدءاً بتعريف المدونات اللغوية مرورًا بلمحة عن طرق الإفادة منها وأنواع المدونات اللغوية ومواصفاتها وإنشائها ومتطلبات التعامل معها، ثم مناقشة تفصيلية لمجالات الإفادة من المدونات اللغوية في فروع البحث اللساني العام والتطبيقي والحاسوبي مع التمثيل لكل فرع.

والموضوع واسع تتطلب مناقشته كتابًا أو أكثر، لكن الغاية من هذا المبحث هو استعراض عام لهذا الحقل من حقول المعرفة لعله يكون تمهيدًا مناسبًا للكتاب الذي بين أيدينا والذي يعتبر رائدًا في هذا المجال.

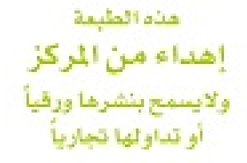
قائمة بمدونات لغوية عربية

أولاً: ترجمة لقائمة لطيفة السليطي المؤرخة في ٢٠ فبراير، ٢٠١٠

اسم المدونة	إنتاج	نوع اللغة	الحجم	الغرض	مصدر المادة النصية
مدونة باك – ولتر العربية ١٩٦٨-٢٠٠٣	تم ولتر	نصوص مكتوبة	٢,٥-٣ بليون كلمة	معجمي	المصادر المنشورة على الشابكة

مُدَوَّنة ليوفان	الجــامعة الكاثوليكية في بلجيكا	نصوص مكتوبة ومنطوقة	٣ مليون منها ٧٠٠,٠٠٠ منطوقة	عمل قاموس هولندي عربي والعكس لمتعلمي اللغة	مصادر من الشابكة، والإذاعة والتلفاز، وكتب دراسة للمرحلة الإبتدائية
مدونة نيوزواير العربية ١٩٩٤	جامعة بنسلفانيا	مكتوبة	٨٠ مليون كلمة	التعليم وتطوير تقنياته	وكالة فرنسا للصحف ووكالة شينخوا وأمة بريس
مدونة كول فرند ١٩٩٥	جامعة بنسلفانيا	محادثات	٦٠ محادثة هاتفية	تطوير تقنيات اللغة	ناطقون باللهجة المصرية
مدونة نيميجن ١٩٦٩		مكتوبة	أكثر من مليوني كلمة	عمل معجم هولندي عربي والعكس لمتعلمي اللغة	مجلات وروايات
كول هوم ١٩٩٧	جامعة بنسلفانيا	محادثات	١٢٠ محادثة هاتفية	التعرف على اللغة من خطوط الهاتف	ناطقون باللهجة المصرية
كلارا ١٩٩٧	جامعة تشالرز براغ	مكتوبة	٥٠ مليون كلمة	عمل معجمي	الدوريات العلمية والكتب ومصادر من الإنترنت من ١٩٩٥ حتى تاريخه
مصر ١٩٩٩	جامعة جون هوبكينز (واشنطن)	مكتوبة	غير محدد	الترجمة الآلية	مدونة ثنائية اللغة للقرآن الكريم

الأخبار الإذاعية وكلام ٢٠٠٠	جامعة بنسلفانيا وإل دي سي	منطوقة	١١٠ برامج إذاعية	التعرف على الكلام	إذاعة صوت أمريكا
مدونة دينار٢٠٠٠	جامعة نيميجن بهولندا وجامعة ليون دو الفرنسية بالتعاون مع سوتيتيل (التونسية)	مكتوبة	١٠٠ مليون كلمة	أغراض معجمية ومعالجة اللغات الطبيعية عامة	غير معروفة
مدونة النهار ٢٠٠١	إي إل آر إيه	مكتوبة	١٤٠ مليون كلمة	أغراض بحثية عامة	صحيفة النهار اللبنانية
مدونة الحياة ٢٠٠٢	إي إل آر إيه	مكتوبة	١٨,٦ مليون كلمة	هندسة اللغة واسترجاع المعلومات	صحيفة الحياة اللبنانية
جيجا ورد العربية	جامعة بنسلفانيا	مكتوبة	٤٠٠ مليون كلمة	معالجة اللغات الطبيعية- استعادة البيانات- نمذجة اللغة	وكالة الأنباء الفرنسية، ووكالتا أنباء النهار والحياة، ووكالة أنباء شينخوا
مدونة متوازية عر/ إنج ٢٠٠٣	جامعة الكويت	مكتوبة	٣ مليون كلمة	تدريس الترجمة والمعجمية	منشورات المجلس الوطني الكويتي



مدونة اللغة العلمية العامة ٢٠٠٤	جامعة مانشستر	مكتوبة	١٠٦ مليون	دراسة المركبات في العربية	http://www.kisr.edu.kw//science
مدونة العربية الفصحى ٢٠٠٤	جامعة مانشستر	مكتوبة	٥ مليون	تحليل معجمي	www.muhad-dith.org and www.alwaraq.com
مدونة متعددة اللغات ٢٠٠٤	جامعة مانشستر	مكتوبة	١٠٧ مليون (منها ١ مليون عربية)	الترجمة	صفحات تقنية المعلومات
مدونة سوتيتل (التونسية)	سوتيتل لتقنية المعلومات (تونس)	مكتوبة	٨ مليون كلمة	معجمي	مواد أكاديمية أدبية ومجلات علمية

المصدر: http://www.comp.leeds.ac.uk/latifa/arabic_corpora.htm

ثانياً: مدونات لغوية إضافية:

يضاف إلى المدونات المذكورة أعلاه المدونات التالية:

اسم المدونة	إنتاج	نوع اللغة	الحجم	الغرض	مصدر المادة النصية
المدونة اللغوية العربية	مدينة الملك عبدالعزيز للعلوم والتقنية (الرياض)	نصوص مكتوبة	٧٣٢٧٨٠٥٠٩ مليون كلمة	لأغراض مختلفة	مصادر متنوعة قديمة وحديثة وتمثل مناطق جغرافية مختلفة

المدونة اللغوية العربية الدولية	مكتبة الإسكندرية	نصوص مكتوبة	متوقع ١٠٠ مليون كلمة	أغراض مختلفة	مصادر مختلفة
الذخيرة النصية الفصحى لجامعة الملك سعود	مها سليمان الربيعة	نصوص مكتوبة	٥٠٦٠٢٤١٢ كلمة	أغراض مختلفة	مصادر تراثية مختلفة
مكنز صخر	شركة صخر	نصوص مكتوبة	متوقع ٥٠٠ مليون كلمة	تطوير برمجيات حاسوبية لغوية	نصوص نثرية معاصرة
مدونة متعلمي العربية	جامعة أريزونا	نصوص مكتوبة		تحليل أخطاء المتعلمين	كتابات دارسين أجانب على مدى ١٥ عاما
مدونة وطن ٢٠٠٤	مراد عباس	نصوص مكتوبة		معالجة اللغات الطبيعية	٥٠٠٠ مقالة من الوطن
مدونة خليج ٢٠٠٤	مراد عباس	نصوص مكتوبة	٤,١ ميقا بايت	معالجة اللغات الطبيعية	حوالي ٢٠٠٠٠ مقالة من الخليج
المدونة العربية -ara biCorpus	جامعة بريغهام يونغ (الأمريكية)	نصوص مكتوبة ومنطوقة	١٧٣,٦٠٠,٠٠٠ كلمة	أغراض مختلفة	صحافة، أدب، حديث، نثر، عامية مصرية ...

مدونة مسكوك (arab-crunch)	طه الزروي في (الجزائر)	نصوص مكتوبة، متلازمات لفظية		بحث لغوي	وكيبيديا العربية ومكتبة ثواب
المدونة المتوازية (إنجليزي-عربي)	حمودة الصالحي	نصوص مكتوبة	٣٧٩٤٦٧٧ كلمة إنجليزية و٣٧٥٥٧٤١ كلمة عربية	بحث لغوي	نصوص مترجمة في الأمم المتحدة

المصادر:

أبو العزم، عبدالغني. اللغة العربية والمعالجة الآلية: برامج صخر نموذجاً:
www.aljabriabed.net/n31_04abualazm.(2).htm

- مدونتا الوطن والخليج، مرادعباس:
http://sites.google.com/site/mouradabbas9/corporal

مدونة «تونسية: مدونة العربية التونسية للباحثتين كارين ماكنيل وفايزة ميلاد: www.tunisiya.org

مدونة العربية arabiCorpus:
http://arabicorpus.byu.edu

مدونة كلمات:
www.lancs.ac.uk/staff/elhaj/corpora.htm

مدونة متعلمي العربية:

http://l2arabiccorpus.cercll.arizona.edu

(لمزيد من المعلومات حول مدونات متعلمي اللغة، ينظر المبحث الثاني من هذا الكتاب)

مدونة مسكوك، في الموقعين:

http://arabcrunch.com/ar/

http://maskouk.sourceforge.net/index.php?content=doc

مدونة نصوص الأمم المتحدة لحمودة الصالحي انظر:

en.wikipedia.org/…/English-Arabic_ Parallel_Corpus_of_United_Nations_Texts

مدينة الملك عبدالعزيز للعلوم والتقنية (الرياض):

http://www.kacstac.org.sa

مها سليمان الربيعة (١٤٣٣هـ) الدليل المرجعي للذخيرة النصية الفصحى لجامعة الملك سعود:

ksucorpus.ksu.edu.sa/ar/

ثالثاً: مدونات عربية موسومة:

لعل من أهم المدونات العربية الموسومة (أي بها تحشية صرفية ونحوية) ما يلي:

(١) مدونة القرآن الكريم، وهي موسومة صرفياً ونحوياً وبها شبكة دلالية للمفاهيم ontology. وهي من إعداد فريق من الباحثين من جامعة ليدز (مجموعة البحث اللغوي)، وهي في نسختها الرابعة في الأول من مايو ٢٠١١. الموقع: http://corpus.quran.com

(٢) مدونة جامعة بنسلفانيا المشكولة والموسومة بأقسام الكلام والمحللة نحوياً، من إعداد باحثين من جامعة بنسلفانيا بمدينة فيلادلفيا الأمريكية برئاسة محمد المعموري، وتسمى Penn Arabic Treebank:

Arabic Treebank: Part 1 v 3.0 (POS with full vocalization + syntactic analysis). Authors: Mohamed Maamouri (project head), Ann Bies, Tim Buckwalter and Hubert Jin.

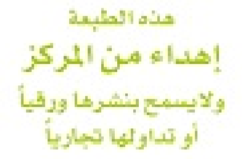
(٣) مدونة جامعة كولومبيا الموسومة صرفياً ونحوياً CATiB، وهي من إنتاج باحثين من جامعة كولومبيا في نيويورك. انظر Nizar Habash، Reem Faraj and Roth, Syntactic Annotation in the Columbia Arabic Treebank

وكذلك : CATiB: The Columbia Arabic Treebank (2009) in: Proceedings of the ICL-IJNLP 2009 Conference Short Parpers.

(٤) مدونة براغ الموسومة نحوياً Prague Arbic Dependency Treebank من إعداد فريق بحث في جامعة تشارلس Charles Universtiy في براغ Prague. وتشتمل المدونة على ١١٣ ألف وخمسمئة كلمة فعلية/ نصية، معظم نصوصها من إصدارات وكالات للأنباء:

http://ufal.mff.cuni.cz/padt/PADT_1.0/docs/index.html

(٥) الذخيرة النصية الفصحى الموسّمة: تذكر منشئة المدونة الباحثة مها الربيعة (في ٧ يونيو ٢٠١٤) عن المدونة اللغوية المذكورة ما يلي: «هي نسخة من الذخيرة النصية الفصحى تحتوي على وسوم توضح الترويسة المعجمية، الجذع، الجذر، الجنس والعدد لكل كلمة



من كلمات الذخيرة النصية. (الدليل المرجعي للذخيرة النصية الفصحى الموسّمة: ksucorpus.ksu.edu.sa/ar/?p=117).

المراجع

المراجع العربية

أبو العزم، عبدالغني « اللغة العربية والمعالجة الآلية: برامج صخر نموذجا» في الموقع: www.aljabriabed.net/n31_04abualazm.(2).htm (تاريخ الدخول ٢٠١٢/٥/١٢).

أبو العزم، عبدالغني (١٩٩٨م). الحاسوب والصناعة المعجماتية، في: مجلة اللسان العربي، العدد ٤٦، ، صص٢٨-٣٩ .

أبو الفتوح، محمد حسين قائمة معجمية بألفاظ القرآن الكريم ودرجات شيوعها. بيروت: مكتبة لبنان.

بايزيد، ليلى (٢٠١٢) «الترابط النصي في المقال الموسوم (فتاة مهمة لسليمان العودة بين الشكل والمضمون» في : دراسات في علم اللغة النصي: مقاربة تطبيقية على مدونة صحيفة الجزيرة (٢٠١٢) الرياض: جامعة الأميرة نورة ، ٢٩٥-٣٦٤.

بن يشو، جيلالي «حوسبة المعجم العربي: الواقع والآفاق» في: مجلة أرتين – المرجع الأول لطلاب الأدب. Art-En.com (تاريخ الدخول ٢٠١٣/٢/٢٢)

البواب، مروان (٢٠٠٩) أثر التقانات الحديثة في تجديد المعجم العربي (تاريخ الإضافة ٢٠٠٩/٣/١٨) www.aluka.net

البواب، مروان (٢٠١٢) محركات البحث في النصوص العربية. من بحوث المؤتمر السابع لمجمع اللغة العربية بدمشق (تاريخ الإضافة ٢٠١٢/٩/٢٣م) www.aluka.net/ literature_language/0/7728/

بوقرة، نعمان (٢٠١٢) «إستراتيجيات الإقناع في الخطاب الصحفي السعودي- دراسة نصية تداولية» في: دراسات في علم اللغة النصي: مقاربة تطبيقية على مدونة صحيفة الجزيرة (٢٠١٢) الرياض: جامعة الأميرة نورة، ١٩-٩٠.

التعابير الاصطلاحية في ضوء النظريات اللسانية الحديثة: دراسة تطبيقية على مدونة صحيفة الجزيرة. الرياض: جامعة الأميرة نورة.

الثبيتي، عبدالمحسن عبيد (٢٠٠٧) «استخدام ذخائر النصوص لاستخلاص المصطلحات المتخصصة» في: الندوة الدولية الأولى عن الحاسب واللغة العربية: الأوراق البحثية، صص ٣١-٣٨.

الثنيان، نوال (٢٠١٢) «الإحالة الضميرية- دراسة نحوية نصية» في: دراسات في علم اللغة النصي: مقاربة تطبيقية على مدونة صحيفة الجزيرة (٢٠١٢) الرياض: جامعة الأميرة نورة، ١٩٣-٢٩٣.

الجلاب، محمد فتحي (١٤٣٢هـ) الاستخلاص الآلي للمحتوى العربي على شبكة الإنترنت بين الواقع والمأمول، في: السجل العلمي لمؤتمر المحتوى العربي، صص ٨٥٣-٨٩.

الحاج صالح، عبدالرحمن (١٩٩٩) «ورقة حول مشروع الذخيرة اللغوية» في: مجلة اللسان العربي، العدد ٤٨.

الحلوة، نوال (٢٠١٢) « أثرالتكرار في التماسك النصي- مقاربة معجمية تطبيقية» في: دراسات في علم اللغة النصي: مقاربة تطبيقية على مدونة صحيفة الجزيرة (٢٠١٢) الرياض: جامعة الأميرة نورة، ٩١-١٩٠.

حمادة، سلوى «المدونات النصية ودور اللغة العربية في التعامل معها». في موقع: www.thomala.com . تاريخ الحصول على الدراسة: ٢٠١٤/٨/١.

حمادة، سلوى (٢٠١١) نحو منهج عربي مُقتَرَح لتصميم المُدَوَّنات اللغويَّة. في موقع: www.globalarabnetwork.com (تاريخ الدخول ٢٠١٤/١/٢).

الحناش، محمد «برنامج لساني-حاسوبي للتعرف الآلي على التعابير المسكوكة في اللغة العربية» في: مجلة التواصل اللساني، ملحق سلسلة الندوات، المجلد ٣، سنة ١٩٩٦، ص ٨٩.

دراسات في علم اللغة النصي: مقاربة تطبيقية على مدونة صحيفة الجزيرة (٢٠١٢) الرياض: جامعة الأميرة نورة.

الربيعة، مها سليمان (١٤٣٣هـ) الدليل المرجعي للذخيرة النصية الفصحى لجامعة الملك سعود. ksucorpus.ksu.edu.sa/ar/ (تاريخ الدخول ٢٠١٣/٩/١).

الزركان، محمد علي (١٩٩٣) «اللسانيات و برمجة اللغة العربية في الحاسوب»، في: وقائع المؤتمر الدولي الأول حول اللغة العربية والتقنيات المتقدمة، الرياض، ١٠-١٩٩٢/٥/١٤، الرياض: مكتبة الملك عبدالعزيز، ١٩٩٣م ، ٥٥-.

سالم، ثناء محمد (٢٠١٢) «الأبعاد التداولية للتعبيرالاصطلاحي» في: التعابير الاصطلاحية في ضوء النظريات اللسانية الحديثة: دراسة تطبيقية على مدونة صحيفة الجزيرة، ١٧-١٢٨، الرياض: جامعة الأميرة نورة.

السجل العلمي: مؤتمر المحتوى العربي في الإنترنت: التحديات والطموحات (٥-١٤٣٢/١١/٧هـ) جامعة الإمام محمد بن سعود الإسلامية، الرياض. المجلد الثاني. الرياض: جامعة الإمام محمد بن سعود الإسلامية.

السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات المنعقدة في مكتبة الملك عبدالعزيز بالرياض (٨-١٢ /١١/ ١٤١٢هـ). الرياض: مكتبة الملك عبدالعزيز .

السجل العلمي لندوة تقنية المعلومات والعلوم الشرعية والعربية المنعقدة في جامعة الإمام محمد بن سعود الإسلامية (١٦-١٤٢٨/٢/١٧هـ) الرياض: جامعة الإمام محمد بن سعود الإسلامية.

السلمان، عبدالملك ومنصور الغامدي وحسن الصبي (١٤٢٨هـ) نظام حاسوبي لرومنة الأسماء العربية. في: الندوة الدولية الأولى عن الحاسب واللغة العربية: الأوراق البحثية، ٢١٥-٢٢٨.

صالح، محمود إسماعيل (٢٠١٢) الحاسوب في البحث اللغوي: لسانيات المدونات أنموذجا. الرياض: كرسي الجزيرة للغة العربية، جامعة الأميرة نورة.

صوالحة، مجدي «العنونة الصرفية والنحوية» في الموقع التالي: www.com.leeds.ac.uk/sawalha/tagset ، تاريخ الوصول ٢٠١١/١/١٤.

صوالحة، مجدي وارك أتول (٢٠١١) «التحليل الصرفي لنصوص العربية الحديثة والكلاسيكية» http://www.comp.leeds.ac.uk/sawalha/sawalha11ICCA.pdf (تاريخ الدخول ٢٠١٤/١١/٢٠)

الصويغ، علي سليمان (١٩٨٧) «كشافات النصوص وتطبيقاتها في نصوص كشافات القرآن والحديث» في: مجلة المكتبات والمعلومات العربية. س ٧، ع٣، ٥-٥٢.

الصيني، محمود إسماعيل (١٤١٢هـ). نحو معجم عربي للتطبيقات الحاسوبية. في: السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات المنعقدة في مكتبة الملك عبدالعزيز بالرياض (٨-١٢ /١١/ ١٤١٢هـ)، ٥١١-٥٢١.

العارف، عبدالرحمن بن حسن «توظيف اللسانيات الحاسوبية في خدمة الدراسات اللغوية العربية: جهود ونتائج» في: مجلة مجمع اللغة العربية الأردني، العدد ٧٣.

عاليم، محمد (١٤٢٨هـ) المعجم العربي في ضوء اللسانيات الحاسوبية. في: الندوة الدولية الأولى عن الحاسب واللغة العربية: الأوراق البحثية، ١٥٧-١٦٦.

عبد العزيز، محمد حسن، ومحمد يونس الحملاوي والمعتز بالله السعيد طه (٢٠٠٨) المعجم الحاسوبي للغة العربية، بحث مقدم في: الاجتماع الثاني لخبراء المعجم الحاسوبي للغة العربية المنعقد بمدينة الملك عبدالعزيز للعلوم والتقنية في الرياض (أبريل ٢٠٠٨).

عبده، داود عطية (١٣٩٩) المفردات الشائعة في اللغة العربية، الرياض: جامعة الملك سعود، ١٣٩٩هـ.

العصيمي، صالح فهد (٢٠١٣) لسانيات المتون وعلوم اللغة، في: مجلة كلية الآداب والعلوم الإنسانية (فاس، المغرب) العدد ١٩، السنة الخامسة والثلاثون، ٣٧-٦٧.

علي، نبيل (١٩٨٧) اللغة العربية والحاسوب: مجلة عالم الفكر، م ١٨، ع ٣، ص ٧٢—

علي، نبيل (١٩٨٨) اللغة العربية والحاسوب. الكويت: مؤسسة تعريب.

علي، نبيل (١٩٩٤) العرب وعصر المعلومات، سلسلة عالم المعرفة، العدد ١٨٤، المجلس الوطني للثقافة والفنون والآداب، الكويت. (ينظر الفصل التاسع من الكتاب).

عمر، أحمد مختار (١٩٨٨) البحث اللغوي عند العرب مع دراسة في التأثير والتأثر، ط٦. القاهرة: عالم الكتب.

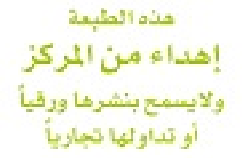
العناتي، وليد وخالد الجبر (٢٠٠٧) دليل الباحث إلى اللسانيات الحاسوبية العربية. عمان: دار جرير للنشر والتوزيع.

غازي، عز الدين (١٤٢٨هـ) قواعد المعطيات المعرفية للمصطلحية العربية: مشروع مقترح. في: الندوة الدولية الأولى عن الحاسب واللغة العربية: الأوراق البحثية، ٢٣-٣٠.

الغامدي، عبدالله شرف و بدرية سليمان الفرهود (١٤٢٨هـ) أداة ويب معتمدة على عملية التحليل الهرمي للحصول على معجم عربي موحد لتقنية المعلومات. في: الندوة الدولية الأولى عن الحاسب واللغة العربية: الأوراق البحثية، ٣٩-٥٠.

الغامدي، منصور محمد «قاعدة بيانات الصوتيات العربية وقراءة الشفاه» www26.brinkster.com/mghamdi (تاريخ الدخول٢٠/١/٢٠١٥).

فايد، وفاء كامل (٢٠٠٠) «بعض مظاهر تغير الصيغ الصرفية في العربية المعاصرة» في: بحوث الندوة الدولية للمعاجم اللغوية العامة والمختصة (المنعقدة في الكويت في ١٤-١٧ مارس، ١٩٩٩).



فايد، وفاء كامل (٢٠٠٧) المتطلبات اللغوية لمعالجة التعابير الاصطلاحية العربية معالجة آلية، في: الندوة الدولية الأولى عن الحاسب واللغة العربية: الأوراق البحثية، ١١-٢٢.

فايد، وفاء كامل (٢٠٠٧) معجم التعابير الاصطلاحية في العربية المعاصرة. القاهرة: د.ن.

الفلاج، نوال (٢٠١٢) «ظاهرة الاقتران الدلالي – دراسة معجمية تطبيقية» في: التعابير الاصطلاحية في ضوء النظريات اللسانية الحديثة: دراسة تطبيقية على مدونة صحيفة الجزيرة، ١٧٧-٢٧١.

الفيفي، عبدالله يحيى (٢٠١٢) المدونات اللغوية لمتعلمي اللغة العربية: نظام لتصنيف وترميز الأخطاء اللغوية. في: السجل العلمبي للمؤتمر الدولي لعلوم وهندسة الحاسوب في اللغة العربية في دورته الثامنة (٢٦-٢٨ ديسمبر، ٢٠١٢) جامعة القاهرة.

القاسمي، علي محمد (٢٠٠٦). لسانيات المدونة الحاسوبية وصناعة المعجم العربي. بحث قدم إلى المؤتمر السنوي الخامس لمجمع اللغة العربية بدمشق، ٢٠-٢٢.

القحطاني، سعد بن هادي «تحليل اللغة العربية بواسطة الحاسب الآلي» مجلة علوم اللغة، م ٥، ع ٣، ٢٢٨- (القاهرة: دار غريب).

الكمار، رأفت (٢٠٠٧) الحاسوب وميكنة اللغة العربية. القاهرة: دار الكتب العلمية للنشر والتوزيع.

لوم، ماري-كلود (٢٠١٢) علم المصطلح: مبادئ وتقنيات، ترجمة د. ريما بركة. بيروت: المنظمة العربية للترجمة.

المالكي، رائدة وهند القحطاني (٢٠١٢) «العلاقات الدلالية للتعبيرات الاصطلاحية في مجال الرياضة» في: التعابير الاصطلاحية في ضوء النظريات اللسانية الحديثة: دراسة تطبيقية على مدونة صحيفة الجزيرة، ١٢٩-١٧٧.

مراياتي، محمد، ويحي مير علم، ومحمد حسن طيان (١٩٩٦) المعجم الحاسوبي: إحصاء الأفعال العربية في المعجم الحاسوبي. بيروت: مكتبة لبنان.

مهدوي، عمر (١٤٣٢هـ) المقاربة الحاسوبية للصرف العربي: قراءة في الحصيلة والآفاق. في: السجل العلمي لمؤتمر المحتوى العربي، ٩٩٩-١٠٢٧.

موسى، علي حلمي وعبدالصبور شاهين(١٩٧٣) دراسة إحصائية لجذور تاج العروس. باستخدام الكمبيوتر. الكويت: جامعة الكويت.

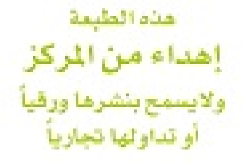
موسى، علي حلمي (١٩٧١) دراسة إحصائية لجذور معجم الصحاح باستخدام الكمبيوتر. الكويت: جامعة الكويت.

موسى، علي حلمي (١٩٧٢) دراسة إحصائية لجذور لسان العرب باستخدام الكمبيوتر. الكويت: جامعة الكويت.

موسى، علي حلمي (٢٠٠١) «حوسبة التراث العربي» محاضرة ألقيت في مجمع اللغة العربي الأردني في ١٧ نيسان (أبريل)، ٢٠٠١. /www.majma. html.1-19-260/28-35-09-10-02-2009/org.jo/majma/index.php (تاريخ الدخول ٢٠١٢/١١/٩).

موسى، علي حلمي (٢٠٠٧) ألفاظ القرآن الكريم: دراسة علمية تكنولوجية. القاهرة: الشركة المتحدة للطباعة والنشر والتوزيع.

ندوة «استخدام الحاسوب في العلوم الشرعية» تحت إشراف مجمع الفقه الإسلامي (٢٤-٢٦ ربيع الآخر، ١٤١١هـ). جدة: البنك الإسلامي للتنمية.



الندوة الدولية الأولى عن الحاسب واللغة العربية: الأوراق البحثية (٢٩/١٠-٢/١١/ ١٤٢٨هـ/ ١٠-٢٠٠٧/١١/١٢). الرياض: مدينة الملك عبدالعزيز للعلوم والتقنية.

النشرتي، مؤمن (١٤٣٢هـ) التحديات التي تواجه محركات البحث في استرجاع المحتوى العربي على الإنترنت: دراسة تحليلية، في: السجل العلمي لمؤتمر المحتوى العربي، ٧١٥-٨١٤.

هلال، يحيى (١٩٩٠) الحاسوب في خدمة الحديث النبوي الشريف. في: ندوة «استخدام الحاسوب في العلوم الشرعية»، ٣٢٩-٣٤٢.

هلال، يحي مختبر المعلوماتية والعلاج الآلي للغة العربية. عرض لنشاط المدرسة المحمدية للمهندسين في الرباط.

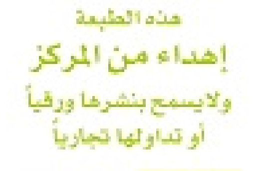
هليل، محمد محمد حلمي (٢٠٠٨) «نحو معجم عربي معاصر» من بحوث الاجتماع الثاني لخبراء المعجم الحاسوبي للغة العربية. الرياض: مدينة الملك عبدالعزير للعلوم والتقنية.

هليل، محمد محمد حلمي وسعد مصلوح وحسان العجمي (٢٠٠٠، تحرير) بحوث الندوة الدولية للمعاجم اللغوية العامة والمختصة .(المنعقدة في الكويت في ١٤-١٧ مارس، ١٩٩٩) الكويت: جامعة الكويت.

المراجع الإنجليزية:

Abdou, Ashraf (2011) Arabic Idioms: A Corpus Based Study. London and New York: Routledge.

Aijmer, K, and Altenberg, B (1991, eds.). English Corpus Linguistics. London and New York: Longman.



Al-ajmi, Hashan. (2004) A new English-Arabic parallel text corpus for lexicographic applications. Lexicos 14 (AFRILEX-reeks/series: 330-326 ,2004 :14.

Al-Ansary, S. (2003). NP-Structure Types in Spoken and Written Modern Standard Arabic (MSA) Corpora. In D. Parkinson & S. Farwaneh (Eds.), Perspectives on Arabic Lin¬guistics XV: Papers from the Fifteenth Annual Symposium on Arabic Linguistics (pp. 180–149). Salt Lake City.

Al-Ansary, Sameh, Nagi, Magdy and Adly, Noha. “Building an International Corpus of Arabic (ICA): Progress of Compilation Stage” at: www.bibalex.org

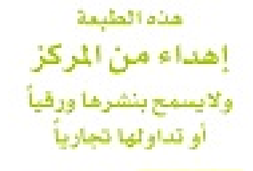
Aliwy, Ahmad H. (2013) Arabic Mophosyntactic Raw Text Part of Speech Tagging. Unpublished Ph.D. dissertation, University of Warsaw.

Al-Muhanna, Amin. (2004). Scientific and technological terms transfer into Arabic: A corpus-based study of Arabic noun+noun and noun+adjective compounds. Ph. D. thesis, UMIST, Manchester.

Al-Saif, Amal and Katja Markert: The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives in Arabic at: www.comp.leeds.ac.uk/markert/Papers/LREC-2010LADTB.pdf

Al-Sulaiti, Latifa and Atwell, Eric. (2006). “The design of a corpus of contemporary Arabic.” International Journal of Corpus Linguistics., vol. 11, pp. 171-135.

Al-Sulaitie, Latifa (2010). Arabic Corpora. At: http://www.comp.leeds.ac.uk./latifa/arabic_corpora.htm



Al-Thubaity, A. O. (2014).A 700M+ Arabic corpus: KACST Arabic corpus design and construction. Language resources and evaluation. DOI 10.1007/s1-9284-014-10579

Al-Thubaity, A. Hend S. Al-Khalifa, Reem Alqifari, and Manal Al-Mazrua, "Proposed Framework for the Evaluation of Standalone Corpora Processing Systems: An Application to Arabic Corpora," Accepted for publication in The Scientific World Journal, Article ID 602745

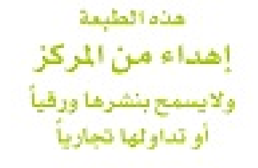
Alzahrani, Rajab Jamaan (2013). "A Corpus-Based Critical Discourse Analysis of the Ideological Representations and Legitimation in the Salafi Discourse in Saudi Arabia (2000-1980)" A Ph. D. dissertation. Lancaster University, U.K.

Arabic Treebank: Part 1 v 3.0 (POS with full vocalization + syntactic analysis). Authors: Mohamed Maamouri (project head), Ann Bies, Tim Buckwalter and Hubert Jin

Atkins, Sue, Jeremy Clear and Nicholas Ostler. (1993) Corpus Design Criteria, in: Literary and Linguistic Computing, 1) 7), Pp.16-1.

Biber, Douglas, Conrad, Susan and Reppen, Randy (1998). Corpus Linguistics: Investigating language structure and use. Cambridge: CUP.

Bowker, Lynne and Pearson, Jennifer (2002). Working with Specialized language: A practical guide to using corpora. London & New York: Routlege.



Buckwalter, Tim and Parkison, Dilworth (2011). A Frequency Dictionary of Arabic: Core Vocabulary for Learners. New Yok and London: Routledge, Taylor, Francis Group

Callies, M. (2008) "Easy to understand but difficult to use? Raising constructions and information packaging in the advanced learner variety" in : Gilquin, G, Papp, S. and Diez-Bedmar, M. (2008.), pp. 226-201.

CATiB: The Columbia Arabic Treebank (2009) in: Proceedings of the ICL-IJNLP 2009 Conference Short Parpers

Cambridge Academic Content Dictionary (2009). Cambridge: Cambridge University Press.

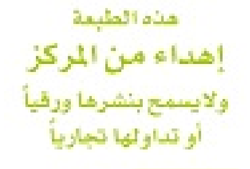
Collins COBUILD English Language Dictionary (5 (2006th Edition. U.K.: Harper Collins Publishers.

Cosme, Ch. (2008) "Participle clauses in learner Englihs" in: Gilquin, G, Papp, S. and Diez-Bedmar, M. (2008.), pp.

Crystal, David (2001) Language and the Internet, Cambridge: Cambridge University Press.

Crystal, David. (1991). "Stylistic profiling" In Aijmer and Altenberg (1991), pp. 238-221.

Dukes, Kais and Nizar Habash "Morphological Annotation of Quranic Arabic" at: http://www.kaisdukes.com/papers/qmorph-lrec2010.pdf



Dukes, Kais, Eric Atwell and Abdul-Baqee Sharif, "Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank" at: http://www.comp.leeds.ac.uk/scams/papers/qsyntax-lrec2010.pdf

Elewa, Abdul-Hamid (2004) "Collocation and Synonymy in Classical Arabic: A Corpus-Based Study" A Ph. D. thesis, University of Manchester.

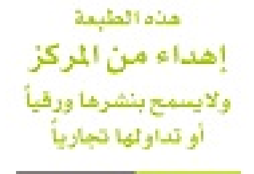
English-Arabic Parallel Corpus of the United Nations Texts. En.wikipedia.org/…/Engish-Arabic_ Parallel_Corpus_of_United_ Nations_Texts.

Facchenetti, Roberta (2007) Corpus Linguistics 25 Years on. (Language and Computers 62) (Language & Computers: Studies in Practical Linguistics). Eiditions Rodopi

Francis, N.W. and Kucera, H. (1979). Brown Corpus Manual: Manual of Information to accompany A Standard Corpus of Present Edited American English for use with Digital Computers, Revised and Amplified. Available at: http://icame.uib./brown/bcm.html#bc2

Garside, R., G. Leech and G. Sampson (eds., 1987). The Computational Analysis of English: A corpus-based approach. London: Longman.

Ghazali, S. & Braham, A. (2001). Dictionary Definitions and Corpus-Based Evidence in Modern Standard Arabic. Arabic NLP Workshop at ACL/EACL,Toulouse,France. (http://www.elsnet.org/arabic2001/ghazali.pdf(



Gilquin, G, Papp, S. and Diez-Bedmar, M. (2008.). Linking up contrastive and learner corpus research. Amsterdam and New York: Rodopi.

Granger, Sylviane (ed.) (1998) Learner English on Computer. London & New York: Addison-Wesley-Longman.

Habash, Nizar, Reem Faraj and Ryan Roth. "Syntactic Annotation in the Columbia Arabic Treebank" at http://www.elda.org/medar-conference/pdf/25.pdf

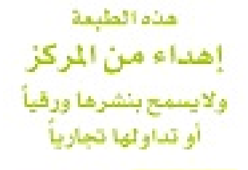
Harley, Trevor (2008) The Psychology of Language: From data to theory, third edition. Hove and New York: Psychology Press (Taylor & Francis Group).

Hassan, Haslina and Nuraihan Mat Daud (2011). Corpus Analysis of Conjunctions: Arabic Learners' Difficulties with Collocations. In: Workshop on Arabic Corpus Linguistics (WACL), 11th 12-th April 2011, Lancaster University, UK. At: http://ucrel.lancs.ac.uk/wacl/slides-HASSAN-DAUD.pdf

Holt, R. (2004) Dialogue on the Internet: Language Civic Identity and computer mediated communication. London: Praeger

Hunston, Susan (2002) Corpora in Applied Linguistics. Cambridge: Cambridge University Press.

James, C. (1998). Errors in Language Learning and Use: Exploring Error Analysis. London: Longman.



Kamal, Eman (2008) The Structure of Arguments in English and Arabic Newspaper Editorials: A Contrastive Study." An unpublished Ph. D. dissertation, King Saud University, Riyadh.

Kennedy, G.D. (1998) An Introduction to Corpus Linguistics. London: Longman.

Kilgarriff, Adam and Grefenstette (2003)"Web as Corpus" in: Computational Linguistics, Vol. 29. at: www.kilgariff.co.uk./Publications/-2003KilgGrefesnstette-WACIntro.pdf

Lawler, J. and H.A. Dry (eds. 1998). Using Computers in Linguistics: A Practical Guide. London and New York: Routledge.

Leech, Geoffrey (2004) Adding Linguistic Annotation in: Wynne (2005).

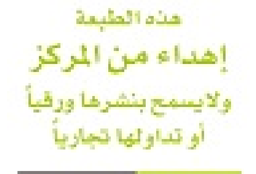
Longman Dictionary of Contemporary English (5 ,(2009th Edition. U.K.: Pearson Longman Publishers.

McEnery, T. & Wilson, A (2001) English Corpus Linguistics, 2nd Ed. Edinburgh: Edinburgh University Press.

McEnry, Tony, Xiao, Richard and Tono, Yukio (2006). Corpus-Based Language Studies: An advanced resource book. London & New York: Routledge.

Meyer, C (2002) English Corpus Linguistics: An Introduction. Cambridge: Cambridge University Press.

MOSES Machine Translation System. At http://www.statmt.org/moses



Myers, G. (2010) Language of Blogs and Wikis. London: Continuum International Publishing Group.

O'Keeffe, Anne, McCarthy, Michael and Carter, Ronald (2007). From Corpus to Classroom: Language Use and Language Teaching. Cambridge: CUP.

Olohan, Maeve (2004). Introducing Corpora in Translation Studies. London: Routledge

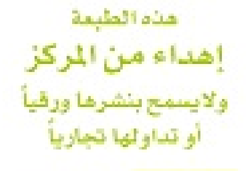
Osborne, John (2008), "Adverb placement in post-intermediate English: a contrastive study of learner corpora" in Gilquin, G, Papp, S. and Diez-Bedmar, M. (1270146 ,(.2008.

Parkinson, D. (1985). Constructing the social context of communication: terms of address in Egyptian Arabic. Berlin/New York/Amsterdam: Mouton de Gruyter.

Parkinson, D. (2003). Future Variability: A Corpus Study of Arabic Future Particles. In D. Parkinson & S. Farwaneh (Eds.), Perspectives on Arabic Linguistics XV: Papers from the Fifteenth Annual Symposium on Arabic Linguistics. Salt Lake City, Amsterdam: John Benjamins Publishers.

Parkinson, D.& Farwaneh, S. (Eds.) (2003). Perspectives on Arabic Linguistics XV: Papers from the Fifteenth Annual Symposium on Arabic Linguistics, Salt Lake City. . Amsterdam: John Benjamins Publishers.

Richards, Jack and Theodore Rodgers (2014). Approaches and Methods in Language Teaching, Third Edition. Cambridge: Cambridge University Press.



Sawalha, M. and Atwell, E. (2008). “Comparative Evaluation of Arabic Language Analysers and Stemmers” in: Coling 2008: Posters and Demonstrations, pp. 110-107.

Scholfied, P. (1995). Quantifying Language: A Researcher’s and Teacher’s Guide in Computing Language Data and Reducing it to Figures. Clevedon, U.K.: Multilingual Matters.

Scott, M. and Ch. Tribble. (2006) Textual Patterns: Key words and corpus analysis in language education. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Second Workshop on Arabic Corpus Linguistics at Lancaster University (July, 2013) See: http://www.comp.leeds.ac.uk/eric/wacl/wacl2proceedings.pdf

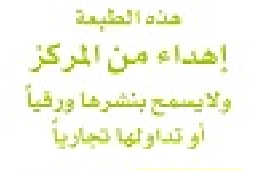
Sieny, Mahmoud (1986) “Tense and Aspect in English and Arabic: Communicative and Functional Equivalence” in : Bulletin of the College of Arts, King Saud University, Vol. 59-41 ,13.

Sinclair, J. (ed., 1987). Looking Up: An account of the COBUILD Project in lexical computing. London and Glasgow: Collins ELT.

Sinclair, J. (1991) Corpus, Concordance and Collocation. Oxford: Oxford University Press

Smith, M. (2009) Online Communication: Linking Technology, Identity and Culture. London: Routledge.

Stubbs, M. (1996) Text and Corpus Analysis: Computer Assisted Studies of Language and Culture. Oxford: Blackwell.



Taweel, Abeer Q., Saidat, Emad M Rafayah, Hussein A., & Saidat, Ahmad M. (2011). Hedging in Political Discourse, in: The Linguistics Journal. June 2011 Volume 5 Issue 196-169 ,1.

Taylor, S. (2003). Comparing Frequencies of Lexical Productions in Arabic Words. In D. Parkinson & S. Farwaneh (Eds.) Perspectives on Arabic Linguistics XV: Papers from the Fifteenth Annual Symposium on Arabic Linguistics (pp. 189–181). Salt Lake City.

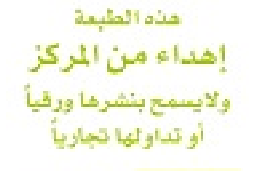
Thomson, P. (2004) Spoken language corpora. In: Wynne (2005) (Ed.) Developing Linguistic Corpora: A Guide to Good Practice. (AHDS Guides to Good Practice) At: http://www.ahds.ac.uk/creating/guides/linguistic-corpora//index.htm

Thorndike, Edward. L. and Lorge, Irving (1972). The Teacher's Word Book of 30,000 Words. New York: Columbia University, Teachers College Press.

van Mol, Mark (2000a). The development of a new learner's dictionary for Modern Standard Arabic: the linguistic corpus approach. In U. Heid, S. Evert, E. Lehmann & C. Rohrer (Eds.), Proceedings of the ninth EURALEX International Congress (pp. 836–831). Stutt¬gart, 12–8 August. (http://www.ilt.kuleuven.ac.be/ilt/arabic/_pdf/stuttgart.pdf)

van Mol, Mark (2000b). Exploring annotated Arabic corpora: preliminary results. (http://www. ilt.kuleuven.ac.be/ilt/arabic/_pdf/tunis.pdf)

van Mol, Mark (2003a). Evolution of MSA, the Case of Some Complementary Particles. In D. Parkinson& S. Farwaneh (Eds.),



Perspectives on Arabic Linguistics XV: Papers from the Fifteenth Annual Symposium on Arabic Linguistics (pp. 147–135). Salt Lake City.

van Mol, Mark (2003b). Variation in Modern Standard Arabic in radio news broadcasts, a syn-chronic descriptive investigation into the use of complementary particles. Bel¬gium: Peeters.

van Mol, M.& Paulussen, H. (2001). AraLat: a relational database for the development of bi-lingual Arabic dictionaries. In S. Lee (Ed.), Proceedings of Asialex 2001, Asian Bilingual¬ism and the Dictionary (pp. 211–206). Seoul, August 2001. (http://www.ilt.kuleuven.ac.be/ilt/arabic/_pdf/asialex.pdf)

West, Michael (1953). A General Service List of English Words: with Semantic Frequencies and a Supplementary word-list for the Writing of Popular Science and Technology, 11th Edition. London: Longman, Green.

Whitelock, P., M.M. Wood, H.L. Somers, R. Johnson and P. Bennett (eds., 1987). Linguistic Theory and Computer Applications. London and New York: Academic Press.

Whitney, Paul (1998.) The Psychology of Language. Boston and New York: Houghton Mififlin Company.

Wynne, M. (2005) (Ed.) Developing Linguistic Corpora: A Guide to Good Practice. (AHDS Guides to Good Practice) At: http://www.ahds.ac.uk/creating/guides/linguistic-corpora//index.htm

المبحث الثاني

مدونات المتعلمين

عبدالله بن يحيى الفيفي

جامعة الإمام محمد بن سعود الإسلامية

الرياض، المملكة العربية السعودية

ayjfaifi@imamu.edu.sa

المقدمة

أثبتت المدونات النصية (Corpora) منذ ظهورها قبل بضعة عقود قدرتها على إعادة الصياغة لمنهجية البحث في كثير من الجوانب اللغوية؛ وذلك لما توفره من إمكانات كبيرة جعلتها أساساً لمادة البحث اللغوي الحديث. هذا التطور الكبير ولَّد عدة أنواع من المدونات النصية، التي تختلف تبعاً لأغراضها وطريقة بنائها، ومن ذلك ما يعرف بمدونات المتعلمين (Learner Corpora). وقد أضحت هذه المدونات إحدى ركائز البحث في مجال اكتساب اللغة وتعليمها؛ لما تكشفه من دلالات على مستويات التعليم اللغوي، ومدى ملاءمة إستراتيجياته؛ ثم إن استخدامها انتشر بشكل كبير فبرزت عدة أنواع فرعية ذات أغراض مختلفة، مع تجاوز استخدامها للبحث في اكتساب وتعليم اللغة إلى مجالات أوسع كتأليف المعاجم، ومعالجة اللغة الطبيعية.

وقد ظهرت في الآونة الأخيرة بعض مدونات المتعلمين العربية، وبدأ استخدامها في مجال تعليم اللغة العربية للناطقين بغيرها؛ فبات من الضروري الحديث عن هذا النوع ووضعه في مكانه الملائم، وفي هذا المبحث نستعرض بعض الأسس النظرية التي وضعها الباحثون مثل تعريف مدونات المتعلمين، وبيان أنواعها، وما الذي يميزها عن غيرها، مع استعراض المدونات العربية منها، لينتقل الحديث بعد ذلك بشيء من التفصيل حول سبعة مجالات تمثل أهم النماذج التي استفادت - ولا زالت تستفيد - من مدونات المتعلمين، وهي ١- أبحاث اكتساب اللغة وتعليمها، ٢- التحليل التقابلي للغة المرحلية، ٣- تحليل الأخطاء بمساعدة الحاسب، ٤- دراسة التطور اللغوي لدى الطلاب، ٥- تأليف المعاجم، ٦- تدريس اللغة وتصميم المواد التعليمية، وأخيراً، ٧- البحث في معالجة اللغة الطبيعية. عند الحديث عن المدونات اللغوية يبرز جلياً موضوع الوسم (١) (Tagging)، وقد اهتمت مدونات المتعلمين بوسم الأخطاء (Error

tagging) أكثر من غيرها؛ لذا خصّصنا لذلك مبحثاً يشرح هذا النوع، ويستعرض الدراسات التي عُنيت بوسم الأخطاء في اللغة العربية، مع الوقوف على بعض النماذج العملية لطريقة وسم الأخطاء. ونختم حديثنا في هذا المبحث بمجموعة من المصادر العلمية لمن أراد القراءة بتوسع حول مدونات المتعلمين.

ما هي مدونات المتعلمين؟

لا فرق من حيث البنية بين مدونات المتعلمين والمدونات العامة، ولذا فإن تعريف قرانجر Granger (٢٠٠٢م) لمدونات المتعلمين - وهي أشهر من عرّف هذا النوع - كان مبنياً على تعريف سينكلير Sinclair (١٩٩٦م) للمدونات العامة. إلا أن ما يميز بين هذين النوعين أمران: الأول أن مصدر البيانات لمدونات المتعلمين ينحصر في تلك المواد التي يحررها متعلمو اللغة دون غيرهم، بينما تجمع نصوص المدونات العامة من مصادر متنوعة، والثاني أن الغاية من بناء مدونات المتعلمين تدور في الغالب حول اكتساب اللغة وتعليمها، في حين تستخدم المدونات العامة لأغراض عدة لا حصر لها. ومن هنا تعرف قرانجر مدونات المتعلمين بأنها:

مجموعة حاسوبية من البيانات النصية الواقعية للغة الثانية أو الأجنبية، والتي تم جمعها وفقاً لمعايير تصميم واضحة، ولغرض محدد في مجال اكتساب اللغة الثانية، أو تعليم اللغة الأجنبية، ويتم وسم هذه النصوص بطريقة معيارية ومتجانسة، مع توثيق أصلها ومصدر الحصول عليها (قرانجر، ٢٠٠٢: ٧)

وحتى يكون التعريف أكثر وضوحاً، فسنشرح بعض مفرداته فيما يلي:

حاسوبية: المقصود بكونها حاسوبية أن تكون محفوظة في الحاسب الآلي بطريقة تسمح بقراءة هذه النصوص آلياً، فالمستندات النصية المدخلة على شكل صور عن طريق الماسح الضوئي مثلاً لا تندرج تحت هذا التعريف.

واقعية: كونها واقعية يعني أن متعلمي اللغة أنتجوها في سياق طبيعي غير مصطنع، ولا بد هنا من التنويه أنه عندما يطلب من الطالب كتابة نص معين بغرض إدراجه في المدونة - أو الحديث عن موضوع معين سواءٌ أكان في مقابلة أم محادثة أم عرض تقديمي أم غير ذلك مع تسجيله صوتياً أو عن طريق الفيديو - فإن لغة الطالب لا تكون واقعية بشكل كامل كما لو كان يحادث أحد زملائه في الفصل مثلاً (قرانجر، ٢٠٠٢م). وبناءً على مقياس نسلهاف Nesselhauf (٢٠٠٤م: ١٢٨) الذي يتألف من أربع درجات: ١- نص طبيعي بشكل كامل، ٢- نص منتج من خلال عملية تعليمية، ٣- نص منتج من خلال مهام تتم مراقبتها والتحكم بها، ٤- وأخيراً برامج نصية ذات سيناريو محدد؛ ترى الباحثة أن مدونات المتعلمين لا تحوي نصوصاً واقعية يمكن تصنيفها تحت الدرجة الأولى، حيث إن منهجية تأليف المدونة تؤثر بشكل واضح في نوعية النصوص المنتجة من الطلاب، فهي في درجة أقل من تلك النصوص الطبيعية الخالصة.

معايير تصميم واضحة: يقصد بهذا تحديد العناصر الأساسية لبناء المدونة، والتي تتبع الهدف من إنشائها غالباً، وهذه العناصر كثيرة نذكر منها على سبيل المثال: ١- تحديد الطلاب الذين ستُضم نصوصُهم للمدونة، ٢- نوع المواد التي ستجمع منهم، ٣- تحديد نطاق المدونة (مؤسسة تعليمية، أو مدينة، أو دولة)، ٤- تحديد حجم المدونة، ٥- مع منهجية جمعها، ٦- ووسمها. ولقد عالج المبحث الثالث من هذا الكتاب موضوع تصميم المدونات بتفصيل أكثر، ولمزيد من المراجع انظر سينكلير (٢٠٠٥)، وقرانجر (٢٠٠٢، و٢٠٠٣ب).

وسم النصوص بطريقة معيارية ومتجانسة: يعد وسم النصوص أحد العناصر المرتبطة بشكل وثيق بالمدونات النصية، ومن هذا المنطلق لا بد أن يستند وسم النصوص في مدونات المتعلمين على إحدى المنهجيات المعيارية المتبعة في مثل هذه العملية، إضافة إلى أهمية تطبيق هذا المنهجية بشكل

متجانس على جميع أجزاء المدونة. (للمزيد حول معنى الوسم - ومثله التحشية - انظر «سابعاً: مصطلحات مهمة في مجال لسانيات المدونات» في المبحث الأول من هذا الكتاب).

توثيق أصلها ومصدر الحصول عليها: يقصد به المعلومات التي تؤخذ مع كل نص بغرض التوثيق وتسمى البيانات الوصفية للمدونة (Corpus metadata)، أو ترويسة المدونة (Corpus header)، ويعرفها برنارد Burnard (٢٠٠٥: ٤٠) بأنها «بيانات حول البيانات»، أي معلومات إضافية حول نصوص المدونة، وهي في الغالب قسمان: الأول بياناتٌ حول مؤلف النص مثل عمره، وجنسه، ولغته الأم، وجنسيته، ومستواه الدراسي أو اللغوي، ومدة دراسته للغة، إلى غير ذلك. والقسم الثاني معلوماتٌ حول النصّ نفسه، مثل نوعه الأدبي (مقال، أو قصة، أو رسالة أو بحث)، وشكله (مكتوب، أو منطوق)، ومكان وتاريخ تأليفه، وعدد كلماته، ونحو ذلك.

بداية مدونات المتعلمين

بدأت مدونات المتعلمين مع ظهور «مدونة لونجمان للمتعلمين» Longman Learners' Corpus (شبكة مدونة لونجمان Longman Corpus Network، ٢٠١٢م) التي بدأ العمل عليها في نهاية ثمانينات القرن الماضي، ومع ظهور مدونات أخرى في ذلك الوقت مثل بنك بيانات اللغة الثانية من مؤسسة العلوم الأوروبية (European Science Foundation Second Language Databank) التي تضم نصوصاً مسجلة لبعض المهاجرين الناطقين بلغات مختلفة، إلا أن عدد العينة في هذه المدونة لم يتجاوز أربعة أشخاص، كما أن البيانات المستخرجة منهم لم تكن واقعية بل كان فيها درجة عالية من التوجيه؛ ولذا اعتُبرت مدونة غير معيارية مع أنها استوفت بقية معايير مدونات المتعلمين كإمكانية قراءة النصوص آلياً (نسلهاف، ٢٠٠٤م). استمر بعد ذلك إنشاء مدونات المتعلمين

على مدى العقدين الماضيين إلى أن فاقت المئة بحسب قائمة قرانجر لمدونات المتعلمين حول العالم (قرانجر، ٢٠١٢م)، إضافة إلى وجود عدد آخر منها لم يدرج بعد في هذه القائمة. وبعد هذا الانتشار لمدونات المتعلمين أصبح الكثير منها مفتوح المصدر، يمكن تنزيل نصوصها والبحث فيها باستخدام أي برنامج لتحليل المدونات، مثل المدونة اللغوية لمتعلمي اللغة العربية (الفيفي وآخرون Alfaifi et al.، ٢٠١٤م)، ومدونة المتعلمين العربية للنصوص المكتوبة (فروانة وتميمي Farwaneh and Tamimi، ٢٠١٢م)، وبعضها متاح للاستخدام من خلال مواقعها الخاصة على شبكة الإنترنت، مثل مدونة ميتشغان الأكاديمية للإنجليزية المنطوقة The Michigan Corpus of Academic Spoken English (سمبسون وآخرون Simpson et al.، ٢٠٠٢م)، والمدونة الروسية لمتعلمي الترجمة Russian Learner Translator Corpus (سوسنينا Sosnina، ٢٠١٤م)، ومنها ما هو تجاري مثل مدونة لونجمان السالفة الذكر، وكذلك مدونة كامبريدج للمتعلمين Cambridge Learner Corpus (٢٠١٢م)، وأيضاً المدونة الدولية لمتعلمي اللغة الإنجليزية International Corpus of Learner English (قرانجر، ٢٠٠٣م ب).

حدود مدونات المتعلمين

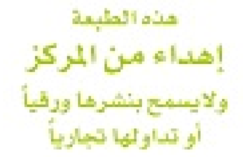
مع اختلاف أنواع متعلمي اللغة فإنه قد يشكل على بعض الدارسين أحياناً ما يمكن أن يدخل تحت مسمى مدونات المتعلمين، فهل يشمل هذا النوعُ النصوصَ التي حررها متعلمو اللغة وإن كانوا من الناطقين بها (الذين يدرسون قواعدها الإملائية والنحوية مثلاً)؟ أو أنها مقصورة على مواد من إنتاج الناطقين بغيرها؟ وللإجابة على هذا السؤال سنستعرض الأنواع الثلاثة التالية:

الأول: مدونات لمتعلمي اللغة باعتبارها لغة ثانية أو أجنبية، وهذا النوع هو المقصود مباشرة بمدونات المتعلمين. ومثال هذا النوع مدونة المتعلمين التمهيدية

للغة العربية (أبو حكيمة وآخرون، ٢٠٠٨م)، وكذلك مدونة متعلمي العربية المكتوبة (فروانة وتميمي، ٢٠١٢م).

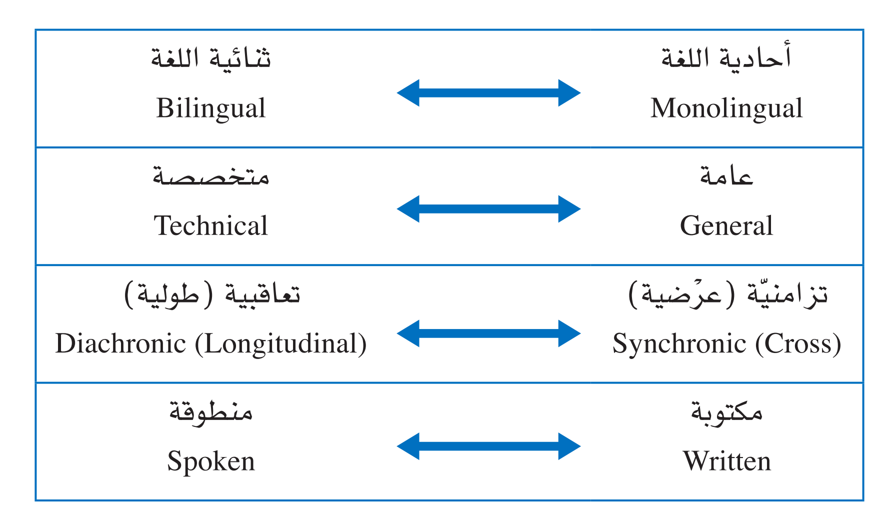
الثاني: مدونات جمعت بين النوع الأول إضافة إلى مدونات متعلمي اللغة الناطقين بها باعتبارها لغتهم الأم؛ وذلك بغرض المقارنة بينهما. وهذا النوع يدخل أيضاً ضمن مدونات المتعلمين؛ وقد وجد الباحث عند دراسته لأكثر من مئة وخمسين مدونة من مدونات المتعلمين (الفيفي، لم ينشر بعد) أن عشرين في المئة منها تندرج تحت هذا النوع. ومثاله المدونة اللغوية لمتعلمي اللغة العربية (الفيفي وآخرون، ٢٠١٤م).

النوع الثالث: مدونات أُفرِدتْ لمتعلمي اللغة الناطقين بها باعتبارها لغتهم الأم، وذلك مثل مدونات المتعلمين العرب الذين يدرسون القواعد الإملائية والصرفية والنحوية للغة العربية، سواءٌ أكان ذلك في مراحل التعليم العام أم في التعليم الجامعي، بغرض دراسة السمات اللغوية لهذه الفئة مثلاً، وذلك مثل مدونة لوفين للمقالات الإنجليزية للناطقين بها (عبدالله ونور Abdullah and Noor، ٢٠١٣)؛ وهذا النوع من المدونات - استناداً على تعريف قرانجر (٢٠٠٢م) - لا يدخل ضمن مدونات المتعلمين التي تركز على اكتساب اللغة الثانية وتعليم اللغة الأجنبية (ثودي Thoday، ٢٠٠٧م)؛ لكن مع ملاحظة ازدياد عدد هذا النوع من المدونات، واشتراكها مع مدونات المتعلمين في كثير من الخصائص مثل البناء والأغراض، بل واشتراكهما أحياناً في مدونة واحدة كما هو الحال في النوع الثاني الآنف الذكر، فإنها أقرب ما تكون إلى مدونات المتعلمين منها إلى أي نوع آخر من المدونات.



تصنيف مدونات المتعلمين

تشير قرانجر (٢٠٠٢م) إلى أن المدونات تصنف في الغالب وفق ثنائيات متقابلة، وسنأتي على أربع منها هي الأقرب إلى مدونات المتعلمين (الشكل ١)، مع ملاحظة أن القائمة اليمنى من هذه الثنائيات هي الغالب على مدونات المتعلمين الحالية.



أحادية اللغة Monolingual	ثنائية اللغة Bilingual
عامة General	متخصصة Technical
تزامنيّة (عرْضية) Synchronic (Cross)	تعاقبية (طولية) Diachronic (Longitudinal)
مكتوبة Written	منطوقة Spoken

الشكل (١) أربع ثنائيات لتصنيف مدونات المتعلمين

إن أغلب مدونات المتعلمين تندرج تحت المدونات أحادية اللغة، مع وجود عدد قليل منها ثنائية اللغة (مترجمة) مثل مدونة سبنس Spence (١٩٩٨م).

أما العموم والتخصيص هنا فيقصد به نوع المواد المضمنة في المدونة، لا نوع المشاركين فيها، والفرق بينهما أن مدونات المتعلمين تعد من المدونات المتخصصة مقارنة بالمدونات العامة؛ حيث تقتصر المشاركة فيها على متعلمي اللغة فقط كما ذكرنا سابقاً في تعريفها، ثم تتفرع بعد ذلك بحسب موادها إلى مدونة ذات لغة عامة، ومدونة ذات لغة خاصة، فالعامة تشمل نصوصاً

في عدة مجالات مختلفة دون تخصيص، بينما الخاصة تركز على النصوص المكتوبة لأغراض محددة، كالسياسة، أو التجارة، أو الدين، على أن يكون كتّابُ هذه النصوص من متعلمي اللغة، ومن هذا النوع الأخير تصنف مدونة المتعلمين «إنديانا» للأعمال (Indiana Business learner corpus)، التي جمعها كونر وآخرون Connor et al. (٢٠٠٢م)، وكانت نصوصها محصورة في الأغراض التجارية.

المدونة التزامنية هي تلك التي يتزامن جمع كل مادة أو جزء منها مع بقية المواد أو الأجزاء الأخرى، أي أنها تجمع تقريباً دفعة واحدة دون فواصل زمنية كبيرة، وقد تسمى كذلك مدونة عرْضية أو مقطعيّة. أما المدونة التعاقبية فتُجمع أجزاؤها على فترات زمنية متعاقبة، وتسمى أيضاً طولية(٢)؛ لأن موادها تُجمع بناء على خط زمني معين. أغلب مدونات المتعلمين من النوع الأول وذلك لصعوبة متابعة عينة محددة من متعلمي اللغة لعدة أشهر وربما لبضع سنوات، إلا أن هذا النوع الأخير مفيد لمراقبة وتحليل التطور الواقع بين كل فترة وأخرى. من المدونات التعاقبية المدونة السويدية لتطوير تعليم اللغة الثانية (The ASU corpus) التي جمعها هامربرغ Hammarberg (٢٠١٠م) لتكون مادة لدراسة التطور اللغوي لكل طالب على حدة، أو بالمقارنة مع بقية الطلاب؛ ومنها كذلك مدونة اللغة المرحلية للمتعلمين الصغار (The Corpus of Young Learner Interlanguage) التي جمعها هوسن Housen (٢٠٠٢م) بإجراء مقابلات مع ستة من متعلمي اللغة الإنجليزية، ثلاثة منهم ناطقون بالفرنسية وثلاثة ناطقون بالهولندية، وذلك لفترة زادت على ثلاث سنوات.

يرى كينيدي Kennedy (١٩٩٨م) أهمية اللغة المنطوقة لغلبة التواصل الشفهي بين الناس على التواصل الكتابي، ولكن عند مقارنة مدونات المتعلمين من حيث المحتوى المكتوب والمنطوق يظهر التفوق الكبير للمدونات ذات المواد المكتوبة، ومع أن ليتش Leech (١٩٩٧م) يقترح أن يكون الجزء المنطوق من

المدونة مساوياً على الأقل للجزء المكتوب، إلا أن المشقة الكبيرة في جمع المواد الصوتية وإدخالها للحاسب يفسر سبب التفاوت الكبير بين هذين النوعين.

ماذا يميز مدونات المتعلمين عن غيرها؟

أشرنا في تعريف مدونات المتعلمين إلى اثنتين من السمات التي تميزها عن أنواع المدونات الأخرى، وسنفرد هنا مبحثاً خاصاً لبيانهما بشيء من التفصيل، مع إضافة سمتين أخريين أقل منهما درجة في التمييز بين مدونات المتعلمين وغيرها من المدونات.

مصدر البيانات: تُجمَع نصوص المدونات العامة من مصادر متنوعة، في حين تقتصر مدونات المتعلمين على المواد التي ينتجها متعلمو اللغة دون غيرهم، من نصوص مكتوبة أو أحاديث منطوقة؛ وقد تتنوع هذه المواد ما بين بحث، أو رسالة، أو قصة، أو مقال، أو كلمة تلقى مشافهة، أو إجابات في مقابلة مع معلم أو زميل دراسة، أو أداءً لأي مهمة لغوية تطلب من الطالب. وتجدر الإشارة هنا إلى أهمية أن تكون المواد المجموعة من الطلاب موحدة، أي تمثل أحد الأنواع السابقة، إما لكامل المدونة أو لكل مدونة فرعية منها، وإلا كانت النتيجة خليطاً من البيانات المختلفة التي لا يمكن الاستفادة منها أو تعميم النتائج عليها (قرانجر، ١٩٩٣م).

الغرض: يمكن ملاحظة أن اكتساب اللغة وتعليمها يمثلان الغرض الرئيس الذي تدور حوله مدونات المتعلمين (ثودي، ٢٠٠٧م)، ومع أن هناك أغراضاً أخرى قد تُستقى من هذا النوع مثل برامج تصحيح الأخطاء آلياً في مجال اللغويات الحاسوبية (انظر مثلاً زغواني وآخرين Zaghouani et al.، ٢٠١٤م)، إلا أن غالبية المدونات الحالية والأبحاث المبنية عليها تعكس الاهتمام الواضح بمجال تعليم اللغة واكتسابها.

الوسم: أشكال الوسم التي يمكن استخدامها في المدونات النصية كثيرة ومتنوعة، إلا أن مدونات المتعلمين تتميز بنوع خاص من الوسم، وهو وسم الأخطاء اللغوية، كونه أحد الأركان الأساسية التي تعتمد عليها بعض الأبحاث القائمة على مدونات المتعلمين. وسنورد بشيء من التفصيل بعض هذه الأبحاث عند حديثنا عن طرق الاستفادة من مدونات المتعلمين، كما سنفرد الحديث عن وسم الأخطاء في جزء آخر من هذا الفصل.

الحجم: مع أهمية حجم المدونة باعتباره أحد العناصر الرئيسة في تمثيل اللغة، إلا أن مدونات المتعلمين ذات حجم صغير في الغالب، وأكثرها يقل حجمه عن المليون كلمة، أما أكبرها فيحوي ما يقارب ٢٥ مليون كلمة مثل مدونة كامبريدج للمتعلمين (٢٠١٢م)، ومدونة جامعة هونج كونج للعلوم والتكنولوجيا (ميلتون ونانديني Milton and Nandini، ١٩٩٤م، وبرافيك Pravec، ٢٠٠٢م)، وكلاهما لمتعلمي اللغة الإنجليزية. وترى قرانجر (٢٠٠٣م ب) أن مدونة مكونة من مئتي ألف كلمة لاستخدامها في مجال اكتساب اللغة الثانية تعد كبيرة إذا ما قارنّاها بالعينات البحثية الصغيرة التي يعتمد عليها الباحثون عادة في هذا المجال؛ لكنها في المقابل تعد صغيرة بالنظر إلى ما هو مستخدم في المجالات اللغوية الأخرى، حيث يعتمد الباحثون على مدونات تتكون من مئات الملايين وربما مليارات الكلمات، خصوصاً أن مثل هذه الأرقام الكبيرة أصبحت معياراً لمقارنة أحجام المدونات بعد أن كانت استثناءً في بداية نشأتها. كما أن هذا الرقم الذي ذكرته قرانجر - ٢٠٠,٠٠٠ كلمة - هو ما اعتمدته في المدونة الدولية لمتعلمي اللغة الإنجليزية كحد أدنى لكل واحدة من المدونات الفرعية التي تمثل مجموعة من الطلاب المتحدثين بلغة أم واحدة (انظر قرانجر، ٢٠٠٣م ب).

مدونات المتعلمين العربية

إن المطّلع على المشاريع القائمة حالياً في مجال مدونات المتعلمين لا يمكن أن تخطئ عينه ما تحظى به اللغات الأجنبية، واللغة الإنجليزية على وجه

الخصوص (الفيفي، لم ينشر بعد)، من نصيب في هذه المدونات؛ فبعد دراسة لأكثر من مئة وخمسين مدونة في هذا المجال، وجد الباحث أن اللغة الإنجليزية تدخل في ٥٢٪ منها، بينما كان نصيب اللغة العربية ٣٪ فقط (خمس مدونات فقط). ولكن عند إمعان النظر في هذه المدونات الخمس نجد أن وتيرة الاهتمام بهذا النوع متسارعة وإن بدأت متأخرة، وهذا ينمُّ عن الحاجة الكبيرة لمدونات المتعلمين لما أثبتته الأبحاث القائمة عليها من فوائد علمية وعملية.

وقبل أن نورد معلومات حول مدونات المتعلمين العربية، فإنه من المستحسن أن نشير – خصوصاً لأولئك المهتمين بجانب التصميم وجمع البيانات لهذا النوع – إلى واحدة من أفضل الأمثلة على مدونات المتعلمين، وهي المدونة الدولية لمتعلمي اللغة الإنجليزية (International Corpus of Learner English)، التي أشرفت على بنائها رائدة البحث في هذا المجال سيلفيان قرانجر من جامعة لوفين ببلجيكا (انظر قرانجر، ٢٠٠٣م ب)، وتعتبر هذه المدونة في نظر كثير من الباحثين مدونة معيارية لما تتمتع به من تصميم دقيق، إضافة إلى المحتوى النصي الذي شارك في جمعه عدد كبير من الباحثين حول العالم، وقد تطورت النسخة الثانية من هذه المدونة لتضم ثلاثة ملايين وسبعمئة ألف كلمة لنصوص حررها متعلمو اللغة الإنجليزية في عدد من الدول، يمثلون ست عشرة لغة أم، ويجري العمل حالياً على النسخة الثالثة منها. أما بالنسبة للغة العربية، فهناك خمس مدونات لمتعلميها تتفاوت من حيث الحجم والخصائص، وفيما يلي نبذة موجزة حول كل منها.

المدونة الأولى: مدونة غازي أبو حكيمة وآخرين

١. اسمها: مدونة المتعلمين التمهيدية للغة العربية – The Pilot Arabic Learner Corpus

٢. حجمها: ٩٠٠٠ كلمة

٣. المواد المشمولة فيها: نصوص مكتوبة

٤. الطلاب المشاركون فيها: متعلمو اللغة العربية باعتبارها لغة أجنبية في الولايات المتحدة الأمريكية

٥. الوسم: غير موسومة

٦. نشر المدونة للباحثين: المدونة لا زالت قيد التطوير وهي غير متاحة للاستخدام العام

١. المرجع: أبو حكيمة وآخرون Abuhakema et al. (٢٠٠٨م)

المدونة الثانية: مدونة سميرة فروانة ومحمد تميمي

١. اسمها: مدونة متعلمي العربية المكتوبة – Written Arabic The L2 Corpus

٢. حجمها: غير موضح، لكنه بحسب إحصاء تقديري للنصوص تضم حوالي ٣٥٠٠٠ كلمة

٣. المواد المشمولة فيها: نصوص مكتوبة

٤. الطلاب المشاركون فيها: متعلمو اللغة العربية باعتبارها لغة أجنبية في الولايات المتحدة الأمريكية

٥. الوسم: غير موسومة

٦. نشر المدونة للباحثين: المدونة متاحة على شكل ملفات PDF فقط، ويمكن تنزيلها من موقع المدونة http://l2arabiccorpus.cercll.arizona.edu

٧. المرجع: فروانة وتميمي (٢٠١٢م) (انظر الشكلين ٢، و٣)

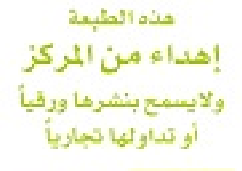
Best Lecture Author 6 L1-Spanish Advanced 4th Year Spontaneous Narrative/Opinion

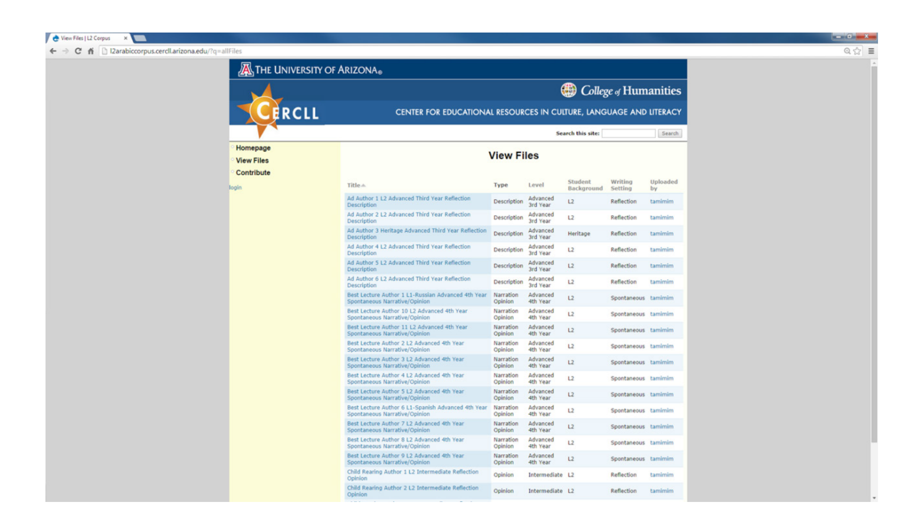
من وجهة نظري كانت محاضرة الدكتور زيدان أفضل محاضرة حضرتها هذا الفصل. تكلم الدكتور زيدان فيها عن ظاهرة تغير المناخ ولا سيما عن مشكلة الاحتباس الحرري، وحسب الاستاذ زيدان سبب هذه المشكلة هي الأنشطة البشرية مثل التلوث البيني، وفي تقديمه استخدم صور ممتعة جداً ولقدم معلومات مهمة ومشوقة. اضافة إلى ذلك يستخدم الدكتور لغة بسيطة وواضحة وخير دليل على ذلك هو أنني فهمت على الأقل 90% من التقديم.

في رأي مشكلة الاحتباس الحرري هي مشكلة جدية ولا بد من أن تتفهمها قبل أن يكون متأخراً. ولذا يجب علينا أن نخطط منهجاً لتعليم الجغرافية وعلم البيئة في المدارس لأن المدارس اليوم تهتم فها أقل مما ينبغي وترتكب خطأ كبيراة إذا لم نُعَلمهم عن ذه المشكلة. كل همي الأن أن نطور جيلاً يهتم بالبيئة والمجتمع.

Best Lecture Author 6 L1-Spanish Adv.4th Year. Spon. Narr/Opin

الشكل (٢) نموذج لملف نصي مع البيانات الوصفية من مدونة المتعلمين العربية للنصوص المكتوبة





الشكل (٣) موقع مدونة المتعلمين العربية للنصوص المكتوبة

المدونة الثالثة: مدونة حسلينا حسان ومحمد فهام محمد غالب

١. اسمها: المدونة الماليزية لمتعلمي العربية – Malaysian Corpus of Arabic Learners

٢. حجمها: حوالي ٨٧٥٠٠ كلمة تقريباً

٣. المواد المشمولة فيها: مقالات وصفية ومقارنة، كتبت باستخدام الحاسب، يصل عددها إلى ٢٥٠ نصاً، متوسط طول النص الواحد ٣٥٠ كلمة

٤. الطلاب المشاركون فيها: متعلمو اللغة العربية الماليزيون في المستوى المتقدم

٥. الوسم: غير موسومة

٦. نشر المدونة للباحثين: هذه المدونة غير متاحة على شبكة الإنترنت، لكن المؤلفة تنوي إدراجها مستقبلاً ضمن مدونة أكاديمية أوسع يمكن

الوصول إليها عن طريق العنوان التالي: http://efolio.iium.edu.my/arabicconcordancer

٧. المرجع: حسان وغالب (٢٠١٣)

المدونة الرابعة: مدونة محمد الكنهل وآخرين

١. اسمها: مدونة تصحيح الأخطاء الإملائية

٢. حجمها: ٦٥٠٠٠ كلمة

٣. المواد المشمولة فيها: نصوص كتبت يدوياً من قبل الطلاب، ثم أدخلت بعد ذلك إلى الحاسب الآلي بواسطة نساخ مستقلين

٤. الطلاب المشاركون فيها: مجموعتان من الطلاب الجامعيين، كل مجموعة من جامعة مختلفة، دون تحديد للتخصصات أو المستويات

٥. الوسم: هناك نسختان لهذه المدونة، الأولى ملفات نصية (txt) وهي غير موسومة، والثانية قاعدة بيانات (مايكروسوفت أكسس - Microsoft Access) تم تصحيح الأخطاء فيها يدوياً.

٦. نشر المدونة للباحثين: يمكن تنزيل نصوص المدونة على شكل ملفات نصية من الرابط التالي: cri.kacst.edu.sa/Resources/TST_DB.rar

٧. المرجع: الكنهل وآخرون Alkanhal et al. (٢٠١٢م) (انظر الشكل ٤)

العنوان:أثر التدخين في الاماكن العامه وعيره نعم: لان التدخين ضار بي صحه المدخن، ولاذلك لضرر التدخين بي الناس وله أمراض كثيرة: سرطان الرئه وعيره من الأمراض. ولذلك منعة بعض الدول الكبير هذه الشائعه التي ثئر على المجتمعات الكبير وتئثر على قوة بناء المجتمع . هل تعتقد أن مجتمع انتشار فيها التدخين تكون أمر سليمه وشبابه صالحين؟ أن من يذخن أمام صديق له قد يكون الضرر الصديق أكثر من المدخن وأثبت في بعض الت جارب على المدخن ومن يجلس أمام المدخن اكتشف أن أثر التدخين أكثر على الصديق الي يجلس دائما مع المدخن. لذلك منعه أكبر الدول التدخين في الأماكن العامه والمكاتب لأن: أكثر الأماكن التي يكون فيها التدخين السالبي، أو (السلبي) كما في الفقر التي قبلها وهو تحارب الكلم بين المدخن والشخص الغير مدخن. ومن سبب منع التدخين في الاماكن العامه والمكاتب، تشبه بعض الناس في الأقتداء فى المدخنين وفعل مايفعلون(شرب الدخان) واذكر قصه بي اختصار يو جد رجل جا ئر مرض السرطان الرئه بسبب مصاحبه لأصدقاء يشربوا الدخان ف أصبح كأن يد خن معهم وهذا أخر منيجلس مع المدخنين لذلك منع الدول السليمه. وفي الختام اذكر ان كل مجتمع سليم يبي على الشباب فإذا فسد الشباب فسد المجتمع ولذلك ننضر إلى التدخين انه (الموت البطيء) لان الشباب أكبر فئه يشربوا التدخين فإذا أصبح الشاب مريض فقد تصبح الأمه في عجز وكل مايضن بي الشباب والناس يمنع في كل مكان. s.w

الشكل (٤) نموذج لملف نصي من مدونة الكنهل

المدونة الخامسة: مدونة عبدالله الفيفي وإريك أتويل

١. اسمها: المدونة اللغوية لمتعلمي اللغة العربية – Arabic Learner Corpus

٢. حجمها: ٢٨٢,٧٣٢ كلمة

٣. المواد المشمولة فيها: نصوص مكتوبة يدوياً، وأخرى باستخدام الحاسب، وكذلك مقابلات صوتية مسجلة، تدور جميعها حول موضوعين، الأول سردي والثاني يناقش الاهتمامات الدراسية

٤. الطلاب المشاركون فيها: متعلمو اللغة العربية الناطقون بغيرها ويمثلون ٤٧٪ من محتوى المدونة، إضافة إلى طلاب ناطقين بالعربية يمثلون ٥٣٪، وهم من كلا الجنسين (الذكور ٦٧٪، والإناث ٣٥٪)، يدرسون في عدد من المدارس الثانوية ومعاهد اللغة وجامعات المملكة العربية السعودية، وتتفاوت أعمارهم بين ١٦ و ٤٢ سنة.

٥. الوسم: المدونة المتاحة للتنزيل حالياً غير موسومة، وستتاح مستقبلاً نسخة تشمل وسم الأخطاء اللغوية. هناك نسخ إضافية من المدونة على بعض مواقع البحث، وهي مبينة بالتفصيل على موقع المدونة المذكور أدناه.

٦. نشر المدونة للباحثين: يمكن تنزيل نصوص المدونة في ملفات نصية (txt)، أو ملفات بلغة الترميز الممتدة (xml)، إضافة إلى الأصول المكتوبة يدوياً في ملفات (pdf)، وكذلك التسجيلات الصوتية في ملفات (mp3) من خلال موقع المدونة على الرابط التالي:
http://www.arabiclearnercorpus.com

٧. المرجع: الفيفي وآخرون (٢٠١٤م) (انظر الشكل ٥، و٦)

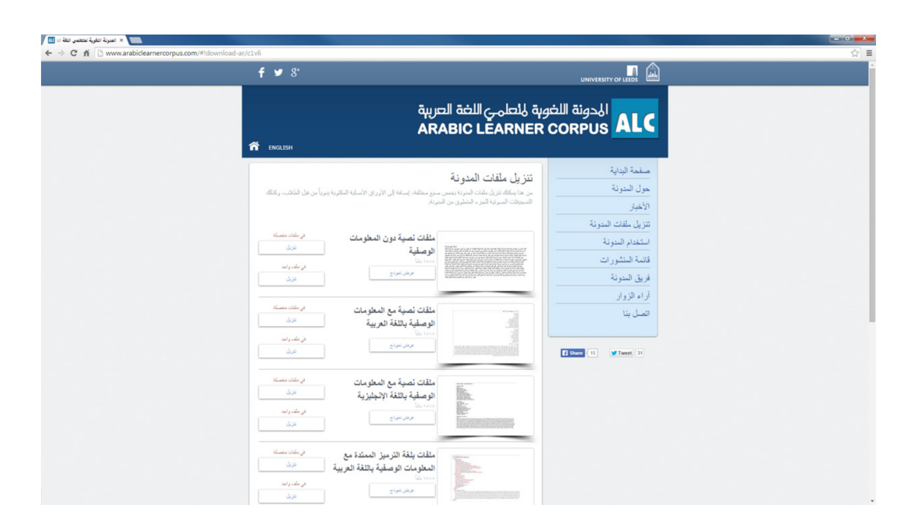
رمز النص: S319_T1_M_Pre_NNAS_W_C

معلومات الطالب
العمر :٢٠
الجنس :ذكر
الجنسية :نيجيري
اللغة الأم :اليوربا
أصالة اللغة :ناطق بغير العربية
عدد اللغات التي يتحدثها :٣
عدد سنوات تعلم اللغة العربية :١٠
عدد السنوات في بلدان عربية :١٠
المرحلة العامة :ما قبل الجامعة
المرحلة الدراسية :برنامج الدبلوم
السنة/المستوى :المستوى الثاني
المؤسسة التعليمية :معهد اللغة بجامعة الإمام

معلومات النص
النوع :سردي
مكان التأليف :في الصف
سنة التأليف :١٤٣٥
دولة التأليف :السعودية
مدينة التأليف :الرياض
محدد بوقت :نعم
استخدام مراجع :لا
استخدام كتاب قواعد :لا
استخدام معجم آحادي :لا
استخدام معجم ثنائي :لا
استخدام مراجع أخرى :لا
الشكل :مكتوب
الوسيط :مكتوب يدويا
الطول :١٧٨ كلمة

عنوان النص :رحلتي إلى لاغوس.
النص :
عرفت أنني سأسافر إلى لاغوس .عاصمة نيجيريا سابقا، في رمضان لأمّ الناس في صلاة التراويح والتهجّد، ولكن الأعمال الكثيرة التي كانت على عاتقي والتي تلزمني التخلص منها تكاد تُنسيني وتمحو السفر عن ذاكرتي .كنت أدرّس بعض الطلاب في إحدى المدارس في مدينة إيوو ولاية أوشن ومما طلب مني إخراج الأسئلة لهم في الإختبار النهائي وفعلت كما طلبوا، وفجأةً جاءت الأوراق للتصحيح .بدأت بتصحيح الأوراق لثلاث مواد درّسها، وقبيل المغرب تذكرت أنني سأسافر إلى لاغوس لأن ذلك اليوم هو التاسع والعشرون للشعبان .ماذا أفعل والأدوات مازالت كثيرة وكأنني لم أصحح منها شيئا أو إضافةً إلى ذلك ما جَهزت للسفر ولا استحمَمْتُ من كثرة الأوراق التي أصحّحَ جاءتني فكرة أن أعطي الأوراق لواحد من الإخوان ثم جهّزت للسفر وتوجهت إلى لاغوس .وصلت إلى المكان الذي أقصده الساعة العاشرة ولكن ا لحمد لله أن الهلال ما طلع في أي مكان ومعنى ذلك أن صلاة التراويح تبدأ في اليوم التالي .بدأنا التراويح في ليلة الثلاثين بعدد غفير من الناس والبهجة والسرور ما اختفيا في أوجههم لأنهم –فضل الله تعالى –شهدوا رمضان عام الفين وثلاثة عشر ميلاديا، أشهر الرحمة والمغفرة وعتق من النار

الشكل (٥) نموذج لملف نصي مع البيانات الوصفية من المدونة اللغوية لمتعلمي اللغة العربية



الشكل (٦) موقع المدونة اللغوية لمتعلمي اللغة العربية

مجالات الإفادة من مدونات المتعلمين

تعد مدونات المتعلمين المصدر الرئيس للبيانات في كثير من الأبحاث اللغوية، حيث قدمت وسائل وأساليب جديدة للبحث أسهمت في الحصول على نتائج أكثر دقة وواقعية، إضافة إلى النتائج العلمية والعملية المستفادة من دراسة وتحليل المفردات والتراكيب اللغوية في مدونات المتعلمين (شابل، ١٤٢٨هـ)؛ وقد قدّم المبحث الأول من هذا الكتاب للقارئ بعض الأمثلة على استخدام المدونات في البحث اللغوي؛ لذا لن نبسط الكلام هنا عن الوسائل والأساليب بقدر ما سنركز حديثنا حول المجالات التي لعبت مدونات المتعلمين فيها دوراً ملموساً مما أدى لتطورها بشكل واضح عما كانت عليه في الماضي.

أولاً: الأبحاث في مجال اكتساب اللغة (Language Acquisition Research)

بناء على الدراستين اللتين أجرتهما قرانجر (١٩٩٨م و ٢٠٠٢م) حول استخدام مدونات المتعلمين في أبحاث اكتساب اللغة فقد وجدت أن هذه الأبحاث

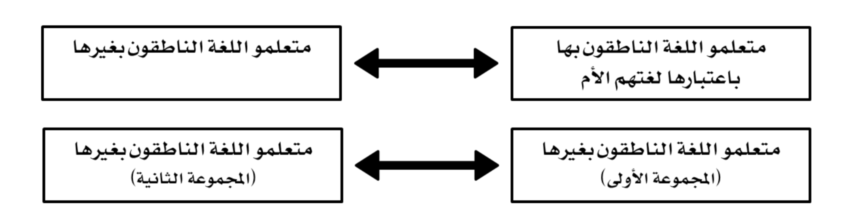
باتت تعتمد بشكل متزايد على هذا النوع من المدونات، وأن هذا التوجه يرجع إلى الأفق الواسع في الإمكانات والأفكار والرؤى التي توفرها مدونات المتعلمين لهذ الحقل الحيوي، الذي يهدف بشكل رئيس إلى فهم آلية اكتساب اللغة الثانية أو الأجنبية؛ إضافة إلى ذلك فإن هذه المدونات - رغم بعض العوامل التي قد تؤثر على نتائجها كما يرى ييب Yip (١٩٩٥م) - تعطي صورة أشمل لبحث جوانب لم يكن بمقدور الدراسات السابقة النظر فيها، ومن ذلك إمكانية دراسة الأشكال اللغوية التي قد يَشُقّ على الطلاب تعلُّمُها، ويصعب عليهم استخدامها، فيختارون تفاديها والالتفاف عليها من خلال أساليب لغوية أخرى، قد تسبب إطناباً واضحاً في غير موضعه.

إضافة إلى ذلك فإن مدونات المتعلمين تمثل أحد أنواع البيانات البحثية التي يمكن من خلالها دراسة اكتساب اللغة، حيث قسّم إليس Ellis (١٩٩٤م) هذه البيانات إلى ثلاثة أنواع، ١- الاستخدام اللغوي: وهي تلك البيانات والنصوص التي تعكس استخدام الطلاب للغة الثانية، سواءٌ أكانت من ناحية الفهم والاستيعاب أم من ناحية الإنتاج؛ ٢- الأحكام اللغوية: وهي الآراء التي يصدرها الطلاب حول اللغة الثانية، كأن يطلب منهم الحكم على جملة لغوية من ناحية التركيب النحوي؛ ٣- التقارير الذاتية: ويقصد بها ما يقدمه الطلاب من معلومات حول إستراتيجياتهم اللغوية عن طريق الاستبانات أو مهام التفكير المسموع ونحوهما. وبالرجوع إلى النوع الأول يقسم إليس «الاستخدام اللغوي» إلى صنفين، يطلق على الأول منهما **بيانات طبيعيّة** في حال لم يتم التحكم في أداء الطلاب عند الحصول عليها، وهذا ما تمثله إلى حد كبير مدونات المتعلمين، ويمثل الصنفَ الثاني **البياناتُ المستنطقة** إن كانت نتيجةً لمهام مقيدة أو أنشطة لغوية ذات رقابة. وفي حين أن الصنف الثاني يقدم معلومات أكثر تركيزاً للجانب المطلوب دراسته فإن مدونات المتعلمين تفتح المجال لدراسة كثير من الجوانب في نفس الوقت (نسلهاف، ٢٠٠٤م)، مع إمكانية دراسة العلاقة

بينها وأثر بعضها على بعض. ويرى إليس (١٩٩٤م) أن البحث الجيد هو ذلك الذي يعتمد على أكثر من نوع من هذه البيانات.

ثانياً: التحليل التقابلي للغة المرحلية (Contrastive Interlanguage Analysis)

التحليل التقابلي (Contrastive Analysis) واحد من المناهج المستخدمة في الأبحاث اللغوية، وتكمن أهميته فيما يكشفه من أوجه التشابه والاختلاف بين اللغات المقارنة؛ وفي مجال مدونات المتعلمين يركز الباحثون على التحليل التقابلي للغة المرحلية من خلال نوعين من المقارنات: النوع الأول بين الناطقين باللغة بوصفها لغتهم الأم ومتعلميها من الناطقين بغيرها، والنوع الثاني بين مجموعتين مختلفتين من متعلمي اللغة الناطقين بغيرها، بحيث تتميزان عن بعضهما في أحد العناصر التي قد تؤثر في عملية التعلم إيجاباً أو سلباً، مثل اللغة الأم للمتعلم، أو العمر، أو المرحلة الدراسية، أو مدة تعلمه للغة، أو عدد اللغات التي يتقنها، أو غير ذلك من العوامل (قرانجر، ١٩٩٨م)، انظر الشكل ٧.



الشّكل (٧) الأنواع الرئيسة للتحليل التقابلي باستخدام مدونات المتعلمين

يهدف النوع الأول إلى الكشف عن الفروق اللغوية بين الناطقين باللغة بوصفها لغتهم الأم مقارنة بمتعلميها الناطقين بغيرها، ويكشف لنا هذا النوع الخصائص المميزة للغة المرحلية للمتعلمين مع تزويد الباحثين بمعلومات دقيقة عنها، فعلى سبيل المثال يمكن معرفة إلى أي مدى يشيع استخدام كلمة أو تركيب

محدد بين كل فريق، ويمكن تبعاً لذلك استخلاص ما يسمى بالاستخدام الزائد، أو الناقص، أو الخاطئ لمفردات وتراكيب اللغة بين متعلميها الناطقين بغيرها؛ وهذا يساعد المختصين من باحثين ومعلمين ومصممي المناهج التعليمية في توجيه الطلاب إلى الاستخدام الصحيح والأمثل للغة، وذلك بتقريبهم ما أمكن إلى أساليب الاستخدام الواقعية للناطقين بها.

أما النوع الثاني من التحليل التقابلي للغة المرحلية فيكون بين مجموعتين مختلفتين من متعلمي اللغة الناطقين بغيرها، حيث تكون المقارنة على أساس أحد العناصر المؤثرة في عملية التعلم، فالمقارنة بين مجموعتين من الطلاب مع اختلاف اللغة الأم لكل منهما (كالمقارنة مثلاً بين مجموعة طلاب إنجليز ومجموعة طلاب صينيين يتعلمون العربية)، يساعدنا في معرفة أثر بعض اللغات على تعلم الطلاب للغة الهدف إيجاباً وسلباً. على سبيل المثال قد يكون تشابه أقسام الضمائر في بعض اللغات عاملاً مساعداً للطالب على إتقانها، بينما يجد غيره صعوبة في فهم الحدود الدقيقة لاستخدامها. ومثل هذا النوع من المقارنات بين مجموعات مختلفة من متعلمي اللغة الناطقين بغيرها، يمكننا من الإجابة على عدة تساؤلات نحو: هل يؤثر عدد اللغات التي يتحدث بها الطالب على مدى قدرته على تعلم اللغة؟ ما مدى أثر العمر على فهم اللغة، واكتساب المفردات، ودقة القواعد، ونحو ذلك؟ هل يؤثر مستوى الطالب لغوياً على شيوع استخدامه لبعض الكلمات أو العبارات؟

ولا بد هنا من التنبيه على دور التصميم الجيد للمدونة؛ ليتمكن الباحثون من إجراء التحليل التقابلي على أكبر عدد ممكن من العوامل، فجودة تصميم المدونة تؤثر بشكل كبير على جودة النتائج المستخرجة منها (سينكلير، ١٩٩١م، وقرانجر، ١٩٩٨م، والسليطي Al-Sulaiti، ٢٠٠٤م).

ثالثاً: تحليل الأخطاء بمساعدة الحاسب (Computer-aided Error Analysis)

بدأ العمل في مجال تحليل الأخطاء اللغوية في نهاية ستينات القرن الماضي – أي قبل بداية وجود مدونات المتعلمين التي بدأت في أواخر ثمانينات القرن نفسه – وقد جُمع كثير من العينات اللغوية للطلاب في ذلك الوقت، لكنها كانت تُستخدم كمستودع للأخطاء فقط، لا على أنها مدونات للمتعلمين، وبالتالي يمكن التفريق بينها وبين مدونات المتعلمين بأن هذه العينات لم تكن تُجمع وفق معايير تصميم واضحة، ولم يكن للحاسب الآلي دور في تحليل محتواها، إضافة إلى أنه كان يتم التخلص منها بمجرد استخراج الأخطاء ودراستها (نسلهاف، ٢٠٠٤م).

ومع استخدام أدوات تحليل المدونات على الحاسب الآلي في الوقت الراهن، أصبح لدى الباحثين إمكانات كبيرة للوصول إلى تحليل أعمق للأخطاء في مدونات المتعلمين، مع الحصول على نتائج أدق وبشكل أسرع. ويلعب وسم الأخطاء الدور الأكبر في هذا الجانب، فكلما كان تصنيف الأخطاء ووسمها يستند على معايير واضحة ودقيقة كانت نتائج البحث والتحليل أكثر جودة وبالتالي أكثر واقعية وفائدة.

ومن الأمثلة على ذلك تلك الدراسة التي أجراها لي Lee (٢٠٠٧م) في جامعة أسكس Essex، حيث قام ببناء مدونة تضم عشرين ألف كلمة، عبارة عن نصوص كتبها طلاب كوريون يدرسون اللغة الإنجليزية، أعمارهم بين ١٦ و١٧ عاماً، وبعد وسم جميع الأخطاء باستخدام أداة خاصة صممها هتشنسون Hutchinson (١٩٩٦م)، وهي أداة تعتمد على تصنيف الأخطاء اللغوية لدنيو وآخرين Dagneaux et al. (١٩٩٦م)، أمكن استنتاج عدد الأخطاء ونسبتها تحت كل فئة من فئات الخطأ العامة والفرعية بسرعة ودقة عاليتين. فعلى سبيل المثال وجد الباحث أن أخطاء القواعد اللغوية كانت الأكثر بنسبة ٣٧,٥٪، تلتها أخطاء علامات الترقيم بنسبة ١٤,٥٪. وفيما يخص التصنيفات الفرعية، وجد أن علامات التعريف والتنكير –

تحت تصنيف القواعد – كانت الأكثر عرضة للخطأ بنسبة ٨, ١٤٪. وهذا النوع من الأبحاث في استنتاج الأخطاء وتحليلها، يستخدم كأساس لكثير من الأبحاث المتقدمة، والأغراض التربوية، فهو يساعد الباحثين على تتبع مستويات الطلاب ومدى تطور قدراتهم اللغوية، إضافة إلى تقصي أوجه القصور في الإستراتيجيات الحاليّة لتعليم اللغة، ومن ثم العمل على تلافيها مستقبلاً، كما يساعد على تصميم كتب، ومواد تعليمية، ومعاجم طلابية أقدر على الوفاء باحتياجات المتعلمين وأهدافهم، وسيرد حديث أكثر عن بعض هذه الجوانب.

رابعاً: دراسة التطور اللغوي لدى الطلاب

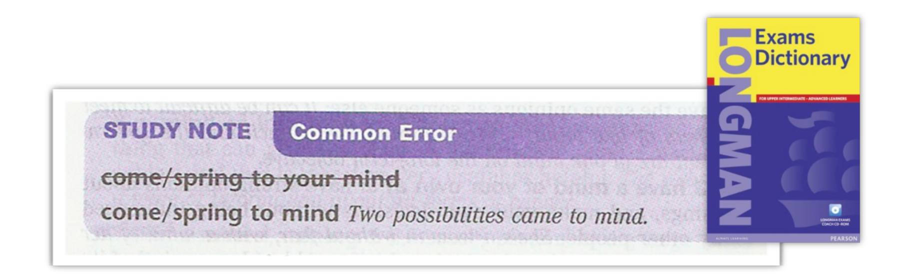
من الأبحاث القائمة على مدونات المتعلمين ما يختص بدراسة مستويات الطلاب وقدراتهم اللغوية وقياس التباين بين كل مستوى وآخر، مع دراسة العوامل المؤثرة في ذلك، ومن الأمثلة على ذلك تلك الدراسة التي قام بها باتيري وكاينز Buttery and Caines (٢٠١٢م) لبحث العلاقة بين مستوى الكفاية اللغوية مقارنة بطول الجمل التي يستخدمها الطلاب من ناحية، وبالتنوع والعدد في استخدام العبارات الظرفية من ناحية أخرى، وقد اعتمدوا في ذلك على وسم وتحليل الأخطاء الذي سبق أن ذكرنا أهميته باعتباره أساساً لمثل هذه الدراسات.

قد يكون هذا المجال هو الأساس الذي بدأ منه ذلك النوع المعروف بمدونات المتعلمين التعاقبية (Diachronic) أو الطولية (Longitudinal)، والتي من أهم أغراضها قياس مدى تطور الطلاب لغوياً سواءٌ أكان ذلك بالاعتماد على نسبة الأخطاء في كل مرحلة، أم من خلال تتبع ما يحدث من تحسن في استخدام بعض المفردات، أو العبارات، أو الأدوات اللغوية، أو غيرها، بالتزامن مع تطور مستوى الكفاية اللغوية لدى الطلاب.

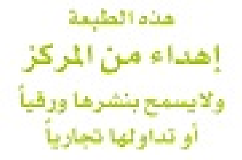
خامساً: تأليف معاجم الطلاب

تعد صناعة المعاجم الموجهة لمتعلمي اللغة – الإنجليزية على وجه التحديد – من أوائل المواد التعليمية التي استفادت من الدراسات المعنية بتحليل مدونات المتعلمين، أما في الوقت الحالي فأغلب معاجم الطلاب تعتمد على هذا النوع من المدونات، خصوصاً المعاجم أحادية اللغة ('Monolingual learners dictionaries). ومن الصور الشائعة لاستخدام مدونات المتعلمين في بناء معاجم الطلاب، الاستفادة من مخرجات تحليل الأخطاء اللغوية على مستوى المفردات أو التراكيب، وتزويد قراء المعجم بتنبيهات حول هذه الأخطاء في المداخل المعجمية المرتبطة بها، مع بيان صور استخدامها الصحيح، ويعد معجم Longman Essential Activator (سمرز Summers، ١٩٩٧م) أول معجم يعتمد على مدونة متعلمين، حيث استُخدِمت «مدونة لونجمان» (شبكة مدونة لونجمان، ٢٠١٢م) في تأليف هذا المعجم (جيلارد وجاسبي Gillard and Gadsby، ١٩٩٨م، وقرانجر، ٢٠٠٣م ب، ونسلهاف، ٢٠٠٤م).

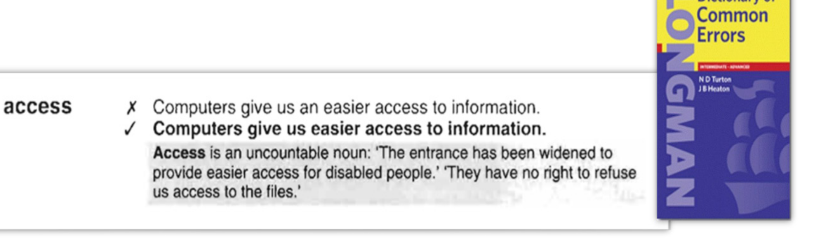
يعرض لنا الشكل ٨ مثالاً لخطأ شائع من «معجم لونجمان للاختبارات» Longman Exams Dictionary (بولن Bullon، ٢٠٠٦م: ٩٦٩)، وقد تم وضعه على شكل ملحوظة في نهاية المدخل المعجمي لكلمة «mind» للتحذير من الأساليب الخاطئة لاستخدام هذه الكلمة والتنبيه على البدائل الصحيحة لها.



الشّكل (٨) مثال لخطأ شائع من «معجم لونجمان للاختبارات»



وقد توسع البحث في تأليف معاجم الطلاب المبنية على مدونات المتعلمين حتى ظهرت معاجم مستقلة للتنبيه على أخطاء المتعلمين وتصويبها مثل «معجم لونجمان للأخطاء الشائعة»(٣) Longman Dictionary of Common Errors (تورتن وهيتن Turton and Heaton، ١٩٩٦: ٥)، انظر الشكل ٩.



الشّكل (٩) مثال لخطأ شائع من «معجم لونجمان للأخطاء الشائعة»

والمعاجم التي تعنى بالتنبيه على الأخطاء الشائعة قد وُجدت بكثرة في اللغة العربية، ومنها «معجم الأخطاء الشائعة» (العدناني، ١٩٨٣م)، و«المعجم الوجيز في الأخطاء الشائعة والإجازات اللغوية» (محمد، ١٤٢٦هـ)، و«معجم تقويم اللغة وتخليصها من الأخطاء الشائعة» (أَمُّون، التاريخ غير معروف)، وهي في الغالب موجهة للناطقين باللغة العربية، ولم تستفد من التطور الحاصل بين حقلي صناعة المعاجم ومدونات المتعلمين.

في مقابل المعاجم هناك بعض الدراسات التي اهتمت بدراسة شيوع الأخطاء لدى متعلمي اللغة، ومن ذلك دراسة العصيلي (١٤٠٥هـ) بعنوان «الأخطاء الشائعة في الكلام لدى طلاب اللغة العربية الناطقين بلغات أخرى»، وكذلك الشافعي وإبراهيم (١٤٠٨هـ) «الأخطاء الشائعة في الهجاء والإملاء بين تلاميذ المرحلة الابتدائية بمنطقة الرياض»، والحمد (١٤١٥هـ) «تحليل أخطاء التعبير الكتابي لدى المستوى المتقدم من دارسي العربية غير الناطقين بها في جامعة الملك سعود» وغيرها من الأبحاث ذات المنهجية الواضحة في

استخراج الأخطاء والحكم على شيوعها، لكنها استخدمت عينات بحثية خاصة عوضاً عن الاعتماد على مدونات المتعلمين، كما أن بناء معجم للطلاب لم يكن جزءاً من أي منها.

سادساً: تدريس اللغة وتصميم المواد التعليمية

من الطرائق المستحدثة في تعليم اللغة، والتي أجريت دراسات كثيرة حول فاعليتها في اكتساب اللغة، تلك التي تُستَخدم فيها المدونات داخل الفصل الدراسي فيما يعرف بالتعليم الموجه بالبيانات (Data-Driven Learning)، والذي يعرفه جونز وكينق Johns and King (١٩٩١م: iii) بأنه:

استخدام البيانات اللغوية التي يتم توليدها داخل الفصل الدراسي، عن طريق استخدام كشاف السياقات (Concordancer) على الحاسب الآلي، وذلك لمساعدة الطلاب على استكشاف العبارات والتراكيب اللغوية، وقياس مدى اطرادها في اللغة المستهدفة، ويشمل ذلك بناء أنشطة وتمارين لغوية اعتماداً على هذه البيانات المستخرجة من كشاف السياقات.

وتمثل لذلك قرانجر وتريبل Granger and Tribble (١٩٩٨م) بتدريب صفي تُستخدم فيه المدونات النصية العامة جنباً إلى جنب مع مدونات المتعلمين، حيث يطلب من الدارسين عمل قائمة بالكلمات التي ترد بعد لفظة محددة، ومقارنة أمثلة الناطقين باللغة (في المدونة العامة) مع أمثلة الناطقين بغيرها (في مدونة المتعلمين).

وسنورد هنا تدريباً مشابهاً لما ذُكر عند قرانجر وتريبل، لكنه مطبق على اللغة العربية، حيث استخدمنا المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية(٤) (الثبيتي Al-Thubaity، ٢٠١٤م) باعتبارها مدونة عامة، والمدونة اللغوية لمتعلمي اللغة العربية (الفيفي وآخرون، ٢٠١٤م) باعتبارها مدونة متعلمين. في هذا التمرين يطلب من المتعلمين ما يلي:

١- إيجاد الكلمات التي تتلو عبارة «بالنسبة» في جميع الأمثلة المعطاة.

٢- فرز هذه الكلمات لإيجاد المفردات التي وردت فقط في مدونة المتعلمين ولم ترد في المدونة العامة، مع محاولة الحكم عليها بالسؤال عن مدى صحة استخدامها كما وردت في الأمثلة، وكيفية تصحيحها، ونحو ذلك.

أولاً: نصوص الناطقين باللغة العربية (مدونة عامة)

وأما العبد فقد يريد الشيء ويكون	بالنسبة	له شراً فيخرج من هذه الجهة عن كونه
وإن كانت مدته أو عمره طويل لكنه	بالنسبة	إلى الزمن بعامة -الزمن المطلق- لا شك أنه
وفق ما تيسر لمؤلفها، والترتيب ينفع المُتَلَقِّي لكن	بالنسبة	لنا سنجري على وفق ما جرى هو عليه
و أصبحت المخاطب المفضل	بالنسبة	للجهات المانحة للمساعدات و المولة للبرامج
فرص الاختيار بين السلع والخدمات سواء	بالنسبة	للمنتجين أو المستهلكين كما إن المعرفة قد تلعب
وهو قلة كلام السلف وعظيم فقههم	بالنسبة	لمن بعدهم . فالصحابة رضي الله عنهم
ومن ثم اختيار القول الراجح في كل صورة . أما	بالنسبة	لتخريج الأحاديث فإن كان الحديث في الصحيحين
الذي هو بالنسبة إلى النحو كأصول الفقه	بالنسبة	إلى الفقه
مما أضعف عزيمة أفراده. أما	بالنسبة	إلى الجيش الإسلامي فقد كان قليل العدد من
وبالعادة يفضل هؤلاء السير على الأقدام . أما	بالنسبة	للحافلات السياحية فإن التحسن السياحي في

ثانياً: نصوص الناطقين بغير اللغة العربية (مدونة متعلمين)

ولهذا أرى النجاح إلى الآن في هذا التخصص	بالنسبة	لي بعيدا، ولكن هدي في وإرادتي وعزمي
مشكلة مع العائلة التي كانت تأويني	بالنسبة	للطعام والأجرة وعدة مشاكل أخرى
اهتماماتي الدراسية هي كثيرة وعديدة ولكن	بالنسبة	إلى التخصص الذي اخترته فهو التخصص العلمي
وذلك أن الحاجة تدعو إليها	بالنسبة	أهل بلدي، معظمهم وأكثرهم محتاجون إلى الدعوة
بعد دراسة اللغة العربية	بالنسبة	للمعهد سألتحق بكلية أصول الدين بإذن الله
هذه هي قصة حياتي وأجمل قصة	بالنسبة	لي فقد قمت بوصفها لك واتمنى أن
وبإذن سيتحقق حلمي الذي اريده، أما	بالنسبة	في اختياري له فليس له سبب
فقد كانت من أكثر المواقف روحانيه	بالنسبة	الي، فلازلت استشعر ذلك الموقف
وايضاً في كلية الشريعة إستفادة كثيرة	بالنسبة	في كليات أخرى
أو يرغب بهذا التخصص.	بالنسبة	أسرتي هم يقولون أي التخصص أريد

الشّكل (١٠) مثال لتدريبات التعليم الموجه بالبيانات

بالإجابة على الأسئلة أعلاه، سيلاحظ الطالب أن الناطقين باللغة العربية في جميع أمثلتهم استخدموا الحرفين «إلى» و«اللام» بعد عبارة «بالنسبة»، بينما استخدم متعلمو اللغة إضافة إلى هذين الحرفين، حرف الجر «في» والاسمين «أهل» و«أسرتي» مما يدفع الطالب للبحث حول صحة هذه التراكيب التي تفرد بها متعلمو اللغة دون الناطقين بها؛ ويظهر لنا من خلال هذا المثال القائم على المدونات النصية – ومنها مدونات المتعلمين – ما يمكن أن يكتسبه الطالب من استخدام لغوي صحيح لبعض العبارات، اعتماداً على بيانات واقعية غير مصطنعة، إضافة لما يكتسبه الطالب عند تفريقه بين الناطقين بالعربية والناطقين بغيرها من الوقوف على بعض الأخطاء المحتملة، والمبنية على بيانات واقعية أيضاً. ولا بد هنا من التنبيه على بعض المحاذير التي ذكرها الباحثون عند استخدام التعليم الموجه بالبيانات (انظر مثلاً تريبل و جونز Tribble and Jones، ١٩٩٧م، وجونز Johns، ١٩٩٤م، وباكرد Packard ، ١٩٩٤م)، ومن ذلك مثلاً الحاجة إلى تدريب الطلاب على استخدام المدونات، وبرامجها، وطرائق البحث فيها، مع الأخذ في الحسبان تفاوت الطلاب في القدرات الحاسوبية واللغوية، ومن المحاذير كذلك أهمية الاختيار الدقيق للأمثلة والسياقات التي تتوافق مع غرض التدريب، وكذلك تصميم التدريبات بما يناسب مستوى الطلاب، فإذا كان طلاب المستوى المتقدم قادرين على تمييز الخطأ في مدونات المتعلمين، فإن غيرهم من المستويات الأقل قد يظنه استخداماً صحيحاً إن لم يجد التوجيه المناسب، وبالتالي تكون المحصلة عكسية بأن يرسخ في ذهنه واحد أو أكثر من الأخطاء الواردة في مفردة أو تركيب لغوي.

أما من ناحية تصميم المواد التعليمية، فقد قارنت بعض الدراسات (مثل لجنق Ljung، ١٩٩١م) بين تعليم اللغة باستخدام مواد تعليمية اعتمدت في تصميمها على المدونات العامة للناطقين باللغة، ومواد تعليمية أخرى مصممة بناء على اجتهاد وحدس المصمم أو المعلم نفسه، وأظهرت النتائج أن مخرجات

النوع الأول كانت أفضل بسبب اعتمادها على أمثلة لغوية حقيقية، وأن نقطة الضعف الكبرى في النوع الثاني تتمثل في إغفال بعض النواحي المهمة لمشاكل الطلاب اللغوية. لكن قرانجر Grangr، (١٩٩٨م) ترى أن المدونات العامة وحدَها لا تشير إلى درجة الصعوبة التي يواجهها متعلمو اللغة، سواءٌ أكان ذلك في المفردات أم في التراكيب، وهنا تأتي أهمية مدونات المتعلمين في تصميم المواد التعليمية، فلا شك أن اعتماد مصممي المواد التعليمية اللغوية على كلا النوعين – المدونات العامة للناطقين باللغة إضافة إلى مدونات المتعلمين الناطقين بغيرها – سيؤدي إلى تطور كبير، مع إمكانية توجيه هذه المواد لتلبي حاجات الطلاب اللغوية بشكل أكبر. ومن الناحية العملية فقد بدأ فعلياً إنتاج بعض الكتب التعليمية المبنية على نتائج أبحاث أجريت على مدونات المتعلمين، وهي نماذج إيجابية مشجعة (قرانجر، ٢٠٠٣م ب).

سابعاً: البحث في معالجة اللغة الطبيعية (Natural Language Processing)

يستفيد مجال معالجة اللغة الطبيعية من مدونات المتعلمين في بعض المجالات المهمة، منها على سبيل المثال التدقيق الإملائي، والتدقيق النحوي، المبنيان على نتائج تحليل الأخطاء، مما يعطي دقة أكبر في اكتشاف الأخطاء الإملائية والنحوية، وكذلك في اقتراح البدائل الصحيحة لها، وترتيبها بحسب أولويتها (انظر نورفيج Norvig، ٢٠٠٧م، و جارفسكي ومارتن Jurafsky and Martin، ٢٠٠٩م، والكنهل وآخرون، ٢٠١٢م)؛ من ذلك أيضاً بناء تطبيقات لتعليم اللغة حاسوبياً، وتستطيع هذه البرامج الإفادة بشكل مباشر من مدونات المتعلمين، لتوليد تدريبات تساعد الطلاب على تجاوز المشاكل الفعلية التي تواجههم؛ وكذلك المساعدة على اختيار نصوص القراءة المناسبة لمستوى الطلاب آلياً، وغير ذلك من التطبيقات الحاسوبية، انظر مثلاً شانق وشانق Chang and Chang (٢٠٠٤م)، والنسخة الإلكترونية من دراسة موريرس Meurers (لما

تطبع بعد) بعنوان: مدونات المتعلمين ومعالجة اللغة الطبيعية (Learner Corpora and Natural Language Processing).

وسم مدونات المتعلمين

كما ذكرنا في بداية الفصل، فإن وسم الأخطاء اللغوية هو أكثر أنواع الوسم شيوعاً في مدونات المتعلمين، حيث تعتمد كثير من الأبحاث في مادتها على نتائج تحليل هذه الأخطاء، كأبحاث اكتساب وتعليم اللغة، وكذلك تأليف المعاجم، وتصميم الكتب التعليمية، ومنها أيضاً بعض تطبيقات اللغويات الحاسوبية. واستناداً على هذا الأساس سنذكر بعض الجوانب المتعلقة بوسم الأخطاء اللغوية في مدونات المتعلمين.

أولاً: ما هو وسم الأخطاء؟

وسم الأخطاء يعني وضع رمز خاص لكل نوع من أنواع الخطأ اللغوي ليسهل تمييزه عن الأنواع الأخرى عند البحث في المدونة وتحليل مادتها، ومن هذه الأنواع مثلاً الخطأ في الهمزة، والخطأ في حروف الكلمة (كزيادة حرف أو إسقاطه أو استبداله بحرف آخر)، ومنها كذلك الخطأ في الإعراب، والخطأ في علامات الترقيم (زيادة علامة ترقيم، أو إسقاطها، أو استبدالها بأخرى خطأً)، وهكذا. وفي الغالب يتم حصر الأخطاء المراد وسمها، ثم تُصنّف بناء على مجالات أو فئات عامة؛ ليسهل فهمُها واستخدامُها من قبل القائمين على عملية الوسم وكذلك التحليل.

ثانياً: الدراسات التي عنيت بتصنيف الأخطاء

منذ عقود مضت ظهرت كثير من الدراسات المهتمة بتحليل الأخطاء اللغوية التي يقع فيها الطلبة في كتاباتهم، ورغم أن هذه الدراسات قد صنَّفتَ هذه الأخطاء إلى عدة أنواع (انظر مثلاً العصيلي، ١٤٠٥ه، والعتيق، ١٤١٢هـ،

والحمد، ١٤١٤هـ، والعقيلي، ١٤١٥هـ)، إلا أن تركيزها ظل محصوراً في العينات البحثية التي تعالجها، متبعة ما يعرف بطريقة التصنيف «من أسفل لأعلى» (Bottom-up approach)، أي أن تكون البداية بالبحث عن الأخطاء في عينة الدراسة وتحليلها، ومن ثم يوضعُ تصنيف لها، ولم يمتد اهتمامها لبناء تصنيفٍ شاملٍ بطريقة البحث «من أعلى لأسفل» (Top-down approach)، حيث يكون التصنيف هو الموضوع أولاً من خلال دراسة أو تصور عام للأخطاء، ومن ثم تتم عملية البحث عن هذه الأخطاء اللغوية وتحليلها في النصوص والمدونات وغيرها. هذه التقسيمات الشاملة للأخطاء لا تقتصر فائدتها على وسم مدونات المتعلمين، بل تتعدى ذلك إلى فوائد أخرى، مثل استخدامها في تعليم اللغة وتصحيح النصوص للطلاب من قِبَل معلميهم، حيث تمنح المعلمَ والطالبَ وضوحاً في تحديد نوع الخطأ بدقة يسهل معها التعرف على أسبابه وطرائق علاجه.

من المحاولات القائمة لبناء تصنيف لأنواع الأخطاء في اللغة العربية، ما قام به غازي أبو حكيمة وآخرون (٢٠٠٨م) من ترجمةٍ لأحد تقسيمات الأخطاء في اللغة الفرنسية (قرانجر، ٢٠٠٣م أ) ومحاولة تطبيقه على اللغة العربية. ويشتمل هذا التصنيف على ثلاث طبقات، الأولى مجالات الخطأ وهي: الشكل، والصرف، والقواعد، والمفردات، والنحو، وازدواجية اللغة، والأسلوب، وعلامات الترقيم، والأخطاء المطبعية؛ أما الطبقة الثانية فحوت فئات الأخطاء الفرعية المندرجة تحت المجالات التسعة السابقة، والطبقة الثالثة جاءت تحت اسم الفئات القواعدية وفيها أربعة أقسام: الاسم، والفعل، وعلامات الترقيم، والرابع لم يسمّه ولعله الحرف، وهذه الطبقة الثالثة عُنيت بالبنية الصرفية أكثر من عنايتها بالأخطاء، انظر أبو حكيمة وآخرين (٢٠٠٨م)، وكذلك الفيفي وأتويل (٢٠١٢م) لمزيد من الإيضاح حول هذا التصنيف.

التصنيف الآخر قام به الباحث كجزء من مشروع المدونة اللغوية لمتعلمي اللغة العربية (الفيفي وآخرون، ٢٠١٣م)، وفيه طبقتان، الأولى تشمل مجالات الخطأ، وقد رُتبتْ تبعاً لتسلسل المستويات اللغوية كما يلي: خطأ إملائي، خطأ صرفي، خطأ نحوي، خطأ دلالي، وبعدها يأتي الخطأ في علامات الترقيم. تحت كل مجال من هذه المجالات الخمسة مجموعة من الفئات الفرعية اعتمد فيها الباحث على عدة دراسات اهتمت بتصنيف الأخطاء وطرق وسمها، وبهذا جمع بين طريقتي البحث السالف ذكرهما، الأولى المسماة «من أسفل لأعلى»، حيث تم بناء تصنيف الأخطاء المقترح اعتماداً على نتائج مجموعة من الدراسات العلمية التي درست عينات كبيرة من أخطاء متعلمي اللغة (انظر الفيفي وأتويل، ٢٠١٢م)، ومن ثم انتقل إلى الطريقة الثانية المسماة «من أعلى لأسفل» من خلال اختبار التصنيف وتطبيقه على عينات من مدونات المتعلمين، ومنها المدونة اللغوية لمتعلمي اللغة العربية. كما قام الباحث بتقويم هذا التصنيف عدة مرات مستقلة لتحسينه وتنقيحه، وفي كل مرة يطبق على عينة مختلفة من النصوص؛ للتأكد من دقته وشموله ووضوح تصنيف الأخطاء فيه. كما قام الباحث بدراسة حول استخدام هذا التصنيف بالمقارنة مع تصنيف أبو حكيمة، وذلك بوسم عينة من النصوص بواسطة باحثَيْنِ مستقلَّيْن، وتم وسم هذه العينة مرتين باستخدام أحد التصنيفين في كل مرة، مع قياس مدى التطابق بين الباحثَين عند استخدام كل واحد من تصنيفات الأخطاء؛ وقد أظهرت النتائج أن التطابق بين المشاركَين عند استخدام التصنيف المقترح كان أكثر منه عند استخدام تصنيف أبو حكيمة، وللاطلاع على نتائج هذه الدراسة انظر الفيفي وآخرين (٢٠١٣م).

ثالثاً: الأخطاء التي توسم

يعتمد ذلك على الغرض من دراسة الأخطاء، فعلى الرغم من استخدام أغلب الباحثين للتصنيفات الموجودة بسبب شمولها لأهم الأخطاء اللغوية، إلا أن هناك من يدرس نقاطاً محددة يعمد فيها إلى استخدام تصنيفه الخاص

الذي يُحقّقُ أهداف بحثه بشكل مباشر، ومن ذلك مثلاً أن يكتفي بوسم الأخطاء التي يجري عليها دراسته مع إضافة بعض التفاصيل عليها. فمن يدرس أخطاء الهمزة على سبيل المثال، قد يتعمق أكثر في تصنيفها لتشمل كل الأخطاء المحتملة، كأن يقسمها مثلاً إلى أخطاء الهمزة في أول الكلمة، والخطأ في الهمزة المتوسطة، والخطأ في الهمزة المتطرفة، وكل واحد منها له أقسامه الفرعية، ففي أول الكلمة قد يتعلق الخطأ بكتابة همزة الوصل أو إسقاط همزة القطع، وفي وسطها قد تُقسّم الأخطاء بناء على الحرف الذي تكتب عليه الهمزة، وهكذا. أما مدونات المتعلمين - موضوع حديثنا في هذا المبحث - فكثير منها يعتمد على تصنيفات الخطأ العامة كتلك التي أوردناها سابقاً (أبو حكيمة وآخرون، ٢٠٠٨م، والفيفي وآخرون، ٢٠١٣م)؛ وسبب ذلك يعود إلى عمومِ الهدف من إنشاء مثل هذه المدونات، فناسب أن يكون وسم الأخطاء فيها عاماً كذلك.

رابعاً: كيفية وسم الأخطاء

يستخدم الباحثون عدة طرق لوسم الأخطاء تعتمد على نوع الملفات المستخدمة لحفظ النصوص (txt أو xml مثلاً)، وكذلك برامج تحليل هذه النصوص، إضافة إلى جدول تصنيف الأخطاء نفسه الذي يفرض أحياناً بعض القيود على طريقة وضع الوسوم. من هذه الطرائق أن تُدرِج وسم الخطأ بعد الكلمة أو العبارة الخاطئة مباشرة، على أن يكون الوسم محصوراً بعلامتين تميزهما عن النص، وهي في الغالب الأقواس المثلثة <>، كما يضيف بعض الباحثين إلى وسم الخطأ الشكل الصحيح للكلمة أو العبارة الخاطئة (انظر الشكل ١١). ويمكن لبعض البرامج المستخدمة في تحليل المدونات - مثل ورد سميث تولز(٥) WordSmith Tools (سكوتّ Scott، ٢٠١٢م)- إدراكُ أن ما بين هذه الأقواس وسوم إضافية على النص وليست من أصله، فلا يتم عرضها مع النص الأصلي، مع تمكين المستخدم من البحث في هذه النصوص وتحليلها بناء على تلك الوسوم.

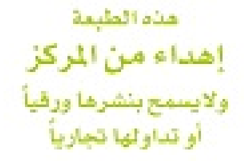
«وما كنتو < OG < كنت < وحيد < OM وحيدًا < في هاذا < OD < هذا < الأمر < PM < الأمر، < بل كانت < XG < كان < معي زملاءي < OH < زملائي < الآخرين < XC < الآخرون <»

الشّكل (١١) مثال لوسم الأخطاء وتصويبها ضمن النص

من طرق وسم الأخطاء كذلك فصل الكلمات بحيث تكون كل كلمة - أو وحدة صرفية أو معجمية - في سطر مستقل، على أن يُجعل وسم الخطأ مقابلاً للكلمة الخاطئة، وتَفصِل بينهما مسافة جدولة (tab)، كذلك يكون التصويب بعد الوسم مع وجود نفس الفاصل (انظر الشكل ١٢)، وتستطيع أداة البحث والتحليل سكتش إنجن(٦) Sketch Engine (كيلقارف Kilgarriff، ٢٠٠٤م) قراءة هذا النوع من التنسيق.

وما		
كنتو	OG	كنت
وحيد	OM	وحيداً
في		
هذا	OD	هذا
الأمر	PM	الأمر،
بل		
كانت	XG	كان
معي		
زملاءي	OH	زملائي
الآخرين	XC	الآخرون

الشّكل (١٢) مثال لوسم الأخطاء وتصويبها على شكل أعمدة



وفيما يلي (الشكل١٣) نموذج تم توليده آلياً بتنسيق XML، من نصوص المدونة اللغوية لمتعلمي اللغة العربية، وذلك بعد وسم وتصحيح الأخطاء فيه.

```xml
<?xml version="1.0"?>
<doc ID="S037_T2_M_Pre_NNAS_W_C">
    <text>
        <title>
            <t id=1>تخصصي</t>
            <t id=2>في</t>
            <t id=3>المستقبل</t>
        </title>
        <p id=1>
            <t id=6>لما</t>
            <t id=7>ننظر</t>
            <t id=8>إلى</t>
            <t id=9>العالم</t>
            <t id=10 ErrTag="XG" ErrForm="الإسلامية " CorrForm="الإسلامي">الإسلامية</t>
            <t id=11>اليوم,</t>
            <t id=12>نرى</t>
            <t id=13>كثيرا</t>
            <t id=14>من</t>
            <t id=15>المشكلات</t>
            <t id=16>وقع</t>
            <t id=17>المسلمون</t>
            <t id=18>فيها,</t>
            <t id=19>ونتساءل</t>
            <t id=20>ما</t>
            <t id=21 ErrTag="PC" ErrForm="الحلّ, " CorrForm="الحلّ؟">الحلّ,</t>
            <t id=22>وماذا</t>
            <t id=23>نفعل</t>
            <t id=24>حتى</t>
            <t id=25>نساهم</t>
            <t id=26>في</t>
            <t id=27>إصلاح</t>
            <t id=28 ErrTag="PC" ErrForm="أمتنا. " CorrForm="أمتنا؟">أمتنا.</t>
            <t id=29>منذ</t>
            <t id=30>أن</t>
            <t id=31>بدأت</t>
            <t id=32 ErrTag="OM" ErrForm="بالاتزام " CorrForm="بالالتزام">بالاتزام</t>
            <t id=33>بالدين</t>
            <t id=34>وتركت</t>
            <t id=35>الجاهلية,</t>
            <t id=36>بدأت</t>
            <t id=37>بالتعرف</t>
            <t id=38>على</t>
            <t id=39>أحوال</t>
            <t id=40>المسلمين</t>
            <t id=41>اليوم</t>
            <t id=42>وفي</t>
            <t id=43 ErrTag="SW" ErrForm="التاريخ, " CorrForm="الماضي،">التاريخ,</t>
            <t id=44>فعرفت</t>
            <t id=45>أنّا</t>
            <t id=46>قد</t>
            <t id=47>كذبنا</t>
```

الشّكل (١٣) ملف بتنسيق XML من المدونة اللغوية لمتعلمي اللغة العربية بعد وسم الأخطاء وتصويبها

وقبل أن نختم حديثنا هنا، نود التنبيه على أن مدونات المتعلمين لم تقتصر على وسم الأخطاء، وإنما فصلنا الحديث حوله لأهميته الخاصة في هذا النوع من المدونات، مع التأكيد على أن مدونات المتعلمين قابلة – كغيرها من المدونات – لوسم الكلمات من الناحية الصرفية، والنحوية، والدلالية، والبلاغية، وغير ذلك من أشكال الوسم التي لا يمكن استقصاؤها هنا، وعلى قدر غنى المدونات من ناحية الوسوم تزداد الفائدة من مادتها. ولنضرب لذلك أمثلة، فعندما تكون الأخطاء موسومة، إضافة إلى وسم الكلمات صرفياً، يمكننا البحث عن الأخطاء في بنية صرفية معينة، مثل أخطاء الهمزة في الأفعال المضارعة. ولو أضفنا لذلك الوسم النحوي لأمكننا البحث مثلاً عن أخطاء المطابقة في الإعراب، أو الجنس، أو العدد، أو التعريف والتنكير، ما بين الأسماء والصفات، وهكذا.

الخاتمة

استعرضنا في هذا المبحث مدونات المتعلمين من عدة جوانب، ووقفنا على مجالات استخدامها، ورأينا ما تزخر به من إمكانات، وما يمكن أن تسهم به في سبيل الرقي بالبحث اللغوي لآفاق أرحب، بدايةً من اكتساب وتعليم اللغة، وتأليف المعاجم، وتصميم المواد التعليمية، وصولاً لاستخدامها في مجال اللغويات الحاسوبية، وتعليم اللغة بمساعدة الحاسب. ومع تعدد هذه الأبحاث والموضوعات فلا زال المجال واسعاً، والطريق ممتدة للاستفادة من مدونات المتعلمين في مجالات جديدة وباستخدام طرائق شتى.

كما أتينا على أهم أنواع الوسم في مدونات المتعلمين، وهو وسم الأخطاء، وأكدنا على أن هذه المدونات يمكنها الاستفادة من أنواع الوسوم الأخرى مما يوسع إمكاناتها ويضاعف طرق البحث فيها. وبما أننا قد أتينا على مجموعة من مدونات المتعلمين العربية في إشارة إلى بداية الاهتمام بها، فإن مما

ينبغي التنبيه عليه في هذا المقام أن العمل في مجال مدونات المتعلمين العربية، ووسمها، والاستفادة منها في مجالات البحث اللغوي – أو غيرها من المجالات – لا يزال في أوله، ويحتاج للكثير من الدراسات ليصل إلى مستويات أفضل، ولعل هذا الفصل قد مهَّد الطريق لذلك، وكوَّن نقطة انطلاق لكثير من الباحثين بإذن الله تعالى.

شكر وعرفان

للزملاء الكرام الذين تفضلوا بمراجعة هذا المبحث، وأبدوا ملاحظات قيمة ساعدتْ على تنقيحه وتهذيبه.

الحواشي

(١) من المصطلحات المستخدمة للدلالة على عملية الوسم «التحشية» أو «الترميز»، وفي هذا الفصل اعتمدت مصطلح الوسم، انظر «قائمة مصطلحات لسانيات المدونات اللغوية» للدكتور محمود إسماعيل صالح:

صالح، محمود إسماعيل: قائمة مصطلحات لسانيات المدونات اللغوية
corpus-/01/2014/http://dr-mahmoud-ismail-saleh.blogspot.co.uk
linguistics.html#more

(٢) يمكن تسميتها كذلك «تراكمية»، أو «تتبعية»، لكني فضلت مصطلح مدونة طولية ليكون مقابلاً مناسباً للمدونة العرضية.

(٣) توجد نسخة معربة من هذا القاموس بعنوان «معجم لونجمان للأخطاء الشائعة في اللغة الإنجليزية (إنجليزي – إنجليزي – عربي)»، نقله إلى العربية نبيل راغب (٢٠٠٧م):

راغب، نبيل: معجم لونجمان للأخطاء الشائعة في اللغة الإنجليزية، الشركة المصرية العالمية للنشر - لونجمان، الجيزة - مصر، ٢٠٠٧م.

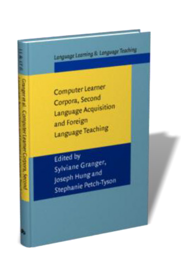
(٤) يمكن الوصول إليها عن طريق الرابط التالي: http://www.kacstac.org.sa

(٥) يمكن تنزيله من الرابط التالي: http://www.lexically.net/wordsmith/

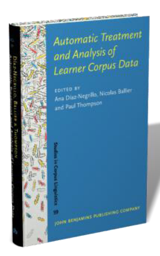
(٦) يمكن الوصول إليها عن طريق الرابط التالي: http://www.sketchengine.co.uk

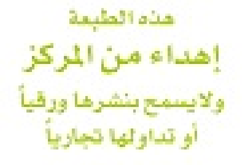
مصادر إضافية حول مدونات المتعلمين

من الملاحظ وجود قصور كبير في المراجع العربية التي تحدثت عن مدونات المتعلمين، ولعل البحث الذي نشر للباحث (الفيفي، ٢٠١٢م) في مؤتمر هندسة وعلوم الحاسب باللغة العربية، من أوائل المراجع العربية التي تناولت هذا النوع من المدونات بشيء من التفصيل، وهو مذكور ضمن قائمة المراجع لهذا الفصل، وسأشير هنا إلى أهم المصادر باللغة الإنجليزية:

مدونات المتعلمين الحاسوبية واكتساب اللغة الثانية وتعليم اللغة الأجنبية **Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching**
يتناول هذا الكتاب مجموعة من الأبحاث مقسمة على ثلاثة محاور رئيسة: • دور مدونات المتعلمين الحاسوبية في اكتساب اللغة الثانية وتعليم اللغة الأجنبية • المناهج القائمة على المدونات في بحث اللغة المرحلية • وسائل تدريس اللغة الأجنبية المبنية على المدونات • الكتاب من إصدار دار النشر الألمانية John Benjamins Publishing Company في ٢٠٠٢م 

عشرون عاماً من البحث في مدونات المتعلمين: نظرة على الماضي للمضي قدماً **Twenty Years of Learner Corpus Research: looking back, moving ahead**
يضم هذا الكتاب أكثر من أربعين بحثاً تمثل السجل العلمي لأول مؤتمر يقام حول مدونات المتعلمين في سبتمبر ٢٠١١ في جامعة لوفاين ببلجيكا، وقد طبعت الأبحاث المقدمة للمؤتمر في هذا الكتاب متسلسلةً دون تصنيف، لكنها شملت عدة موضوعات مثل تصميم مدونات المتعلمين، وجمع بياناتها، ووسمها، وكذلك تحليلها صوتياً، ونحوياً، ودلالياً، وبلاغياً، وغيرها من الجوانب التي تعكس الاهتمام بهذا النوع من المدونات. 

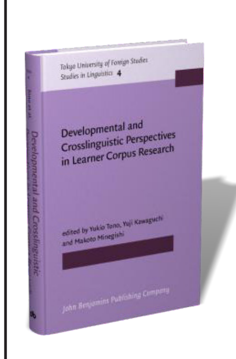
التحليل والمعالجة الآلية للبيانات في مدونات المتعلمين **Automatic Treatment and Analysis of Learner Corpus Data**
يتناول هذا الكتاب مجموعة من جوانب البحث المتعلقة بمدونات المتعلمين، وقد ضم - بعد مقدمة حول مستقبل هذا النوع من المدونات - مجموعة من الأبحاث توزعت على ثلاثة فصول: الأول: جمع مدونات المتعلمين، ووسمها، وتبادل موادها الثاني: وسائل آلية للتعرف على خصائص لغة المتعلمين الثالث: تحليل البيانات في مدونات المتعلمين 



رؤى لغوية تطويرية حول أبحاث مدونات المتعلمين

Developmental and Crosslinguistic Perspectives in Learner Corpus Research

يشتمل هذا الكتاب على فصلين، يركز الفصل الأول على مشروع المدونة النصية الدولية للغة المرحلية (The International Corpus of Crosslinguistic Interlanguage)، وما أجري عليها من أبحاث، بينما يضم الفصل الثاني مجموعة متنوعة من مدونات المتعلمين، ذات البيانات المنطوقة (Spoken learner corpora)، وما تم عليها من أبحاث تتعلق ببنائها، ووسمها، وتحليلها.



المجلة الدولية لأبحاث مدونات المتعلمين

International Journal of Learner Corpus Research



وهي دورية حديثة يبدأ صدور العدد الأول منها مطلع عام ٢٠١٥م من دار النشر الألمانية John Benjamins Publishing Company. تهتم هذه المجلة بنشر الأبحاث التي تتناول بناء مدونات المتعلمين، ووسمها، وتحليلها، واستخدامها في أي مجال من مجالات البحث اللغوي. ورابطها على الشبكة العنكبوتية:

https://benjamins.com/#catalog/journals/ijlcr/main

المراجع

المراجع العربية

أَمُّون، هلا: معجم تقويم اللغة وتخليصها من الأخطاء الشائعة، دار القلم للطباعة والنشر والتوزيع، بيروت، تاريخ الطبع غير معروف.

حسان، حسلينا و غالب، محمد فهام محمد: مشروع جمع المدونات النصية الخاصة بالنصوص الأكاديمية في اللغة العربية، مجلة مجمع اللغة العربية الأردني، العدد الخامس والثمانون، ٢٠١٣م، ٥٧-٧٦.

الحمد، محمد ماجد: تحليل أخطاء التعبير الكتابي لدى المستوى المتقدم من دارسي العربية غير الناطقين بها في جامعة الملك سعود، رسالة علمية غير منشورة، جامعة الإمام محمد بن سعود الإسلامية، الرياض ، ١٤١٤هـ.

شابل، كارول: تطبيقات الحاسب الآلي في اكتساب اللغة الثانية: أسس للتعليم والقياس والبحث العلمي، ترجمة سعد بن علي وهف القحطاني، جامعة الملك سعود، الرياض، ١٤٢٨هـ.

الشافعي، إبراهيم محمد و إبراهيم، عبدالحميد صفوت: الأخطاء الشائعة في الهجاء والإملاء بين تلاميذ المرحلة الابتدائية بمنطقة الرياض، مركز بحوث كلية التربية في جامعة الملك سعود، ١٤٠٨هـ.

العتيق، زايد مهلهل: تحليل الأخطاء الدلالية لدى دارسي اللغة العربية من غير الناطقين بها في مادة التعبير الكتابي، رسالة علمية غير منشورة، جامعة الإمام محمد بن سعود الإسلامية، الرياض، ١٤١٢هـ.

العدناني، محمد: معجم الأخطاء الشائعة، مكتبة لبنان، بيروت، الطبعة الثانية، ١٩٨٣م.

العصيلي، عبدالعزيز إبراهيم: الأخطاء الشائعة في الكلام لدى طلاب اللغة العربية الناطقين بلغات أخرى: دراسة وصفية تحليلية، رسالة علمية غير منشورة، جامعة الإمام محمد بن سعود الإسلامية، الرياض، ١٤٠٥هـ.

العقيلي، عبدالمحسن سالم: تحليل الأخطاء في بعض أنماط الجملة الفعلية للغة العربية في الأداء الكتابي لدى دارسي المستوى المتقدم، رسالة علمية غير منشورة، جامعة الإمام محمد بن سعود الإسلامية، الرياض، ١٤١٥هـ.

الفيفي، عبدالله وأتويل، إيريك: المدونات اللغوية لمتعلمي اللغة العربية: نظام لتصنيف وترميز الأخطاء اللغوية، المؤتمر الدولي الثامن لهندسة وعلوم الحاسب الآلي باللغة العربية، ٢٠١٢م، القاهرة، مصر.

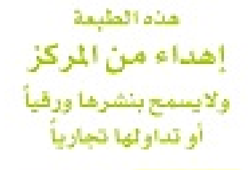
محمد، جُودة مبروك: المعجم الوجيز في الأخطاء الشائعة والإجازات اللغوية، مكتبة الآداب، القاهرة، ١٤٢٦هـ - ٢٠٠٥م

المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية: مدينة الملك عبدالعزيز للعلوم والتقنية، ٢٠١٣م، http://www.kacstac.org.sa

المراجع الأجنبية

Abdullah, Shazila and Noor, Noorzan Mohd: Contrastive Analysis of the Use of Lexical Verbs and Verb-noun Collocations in Two Learner Corpora: WECMEL vs. LOCNESS. In Ishikawa, Shin (Ed.), *Learner corpus studies in Asia and the world.* (Vol. 1), Papers from LCSAW2013. 2013, pp. 139–160.

Abuhakema, G., Faraj, R., Feldman, A. and Fitzpatrick, E: Annotating an Arabic Learner Corpus for Error. In: proceeding of *the*



International Conference on Language Resources and Evaluation, LREC 26 May - 1 June 2008. Marrakech, Morocco, 2008.

Alfaifi, Abdullah, Atwell, Eric and Abuhakema, Ghazi: Error Annotation of the Arabic Learner Corpus: A New Error Tagset. In: *Language Processing and Knowledge in the Web, Lecture Notes in Computer Science. 25th International Conference, GSCL 2013, 25-27 September 2013. Darmstadt, Germany, Springer,* 14 - 22 (9), *2013.*

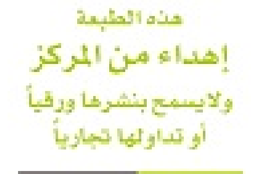
Alfaifi, Abdullah, Atwell, Eric and Hedaya, Ibraheem: Arabic Learner Corpus (ALC) v2: A New Written and Spoken Corpus of Arabic Learners. In the proceedings of the *Learner Corpus Studies in Asia and the World (LCSAW 2014), 31 May - 01 Jun 2014,* Kobe, Japan. <http://www.arabiclearnercorpus.com>, 2014.

Alkanhal, Mohamed, Al-Badrashiny, Mohamed, Alghamdi, Mansour and Al-Qabbany, Abdulaziz: Automatic Stochastic Arabic Spelling Correction With Emphasis on Space Insertions and Deletions. *IEEE Transactions on Audio, Speech, and Language Processing,* 20(7), 2012, pp. 2111–2122.

Al-Sulaiti, Latifa: *Designing and Developing a Corpus of Contemporary Arabic*. Unpublished thesis, University of Leeds, 2004.

Al-Thubaity, Abdulmohsen. O.: A 700M+ Arabic corpus: KACST Arabic corpus design and construction. *Language resources and evaluation*. 2014. DOI 10.1007/s10579-014-9284-1.

Bullon, Stephen: *Longman Exams Dictionary.* Pearson Longman. England, 2006.



Burnard, Lou. Metadata for corpus work. In M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford, Oxbow Books, 2005, pp. 30-46.

Cambridge Learner Corpus: *Cambridge University Press*. 2012, Retrieved from: http://www.cambridge.org/gb/elt/catalogue/subject/custom/item3646603/Cambridge-English-Corpus-Cambridge-Learner-Corpus/?site_locale=en_GB

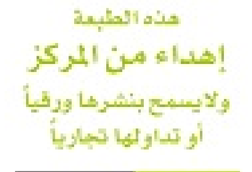
Chang, Jason S. and Chang, Yu-Chia: *Computer Assisted Language Learning Based on Corpora and Natural Language Processing: The Experience of Project CANDLE*. IWLeL 2004, Waseda University, Tokyo. 2004, Retrieved from: https://dspace.wul.waseda.ac.jp/dspace/bitstream/2065/1391/1/02.pdf

Connor, Ulla, Precht, Kristen and Upton, Thomas: Business English: learner data from Belgium, Finland and the U.S. In S. Granger, J. Hung and S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam & Philadelphia, Benjamins, 2002, pp. 175-194.

Dagneaux, Estelle, Denness, Sharon., Granger, Sylviane, and Meunier, Fanny: *Error tagging manual*. version 1.1, 1996.

Ellis, Rod: The Study of Second Language Acquisition. Oxford University Press, Oxford, 1994.

Farwaneh, Samira and Tamimi, Mohammed: *Arabic Learners Written Corpus: A Resource for Research and Learning*. 2012, Retrieved from the University of Arizona, the Center for Educational Resources



in Culture, Language and Literacy Web site: http://l2arabiccorpus.cercll.arizona.edu/?q=homepage

Gillard, Patrick and Gadsby, Adam: Using a learners› corpus in compiling ELT dictionaries. In Granger (Ed.) *Learner English on Computer.* Longman, London and New York, 1998, pp. 159-171.

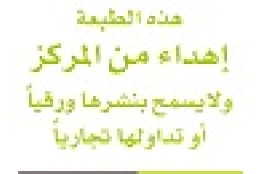
Granger, Sylviane and Tribble, Chris: Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning. In Granger (Ed.) *Learner English on Computer.* Longman, London and New York, 1998, pp. 159-171.

Granger, Sylviane: A Bird›s-eye View of Computer Learner Corpus Research. In S. Granger, J. Hung and S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam & Philadelphia, Benjamins, 2002, pp. 3-33.

Granger, Sylviane: Error-tagged learner corpora and CALL: a promising synergy. *CALICO Journal,* 20(3), 2003 أ, pp. 465-480.

Granger, Sylviane: *Learner corpora around the world*. 2012, Retrieved from the Université Catholique de Louvain, Centre for English Corpus Linguistics Web site: http://www.uclouvain.be/en-cecl-lcworld.html

Granger, Sylviane: The computer learner corpus: a versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on Computer.* London/New York, Addison Wesley Longman, 1998, pp. 3 - 18.



Granger, Sylviane: The International Corpus of Learner English. In J. Aarts, P. de Haan and N. Oostdijk (Eds.), *English Language Corpora: Design, Analysis and Exploitation*. Amsterdam, Rodopi, 1993, pp. 57-69.

Granger, Sylviane: The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. *TESOL Quarterly*, 37(3), 2003 ب, pp. 538-546.

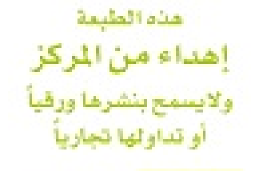
Hammarberg, Björn: *Introduction to the ASU Corpus, a Longitudinal Oral and Written Text Corpus of Adult Learners' Swedish with a Corresponding Part from Native Swedes*. Stockholm University, Department of Linguistics, 2010.

Housen, Alex: A corpus-based study of the L2-acquisition of the English verb system. In: Granger, S., Hung, J. and Petch-Tyson, S. (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam, John Benjamins, 2002, pp. 77-116.

Hutchinson, John: *UCL Error Editor.* Louvain-la-Neuve: Centre for English Corpus Linguistics, Université Catholique de Louvain, 1996.

Johns, Tim and King, Philip: Classroom Concordancing, *ELR-Journal*. 4, 1991, pp. 1-13.

Johns, Tim: From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin (Ed.),



Perspectives on Pedagogical Grammar. Cambridge, Cambridge University Press, 1994, pp. 293-313.

Jurafsky, Daniel, and Martin, James: *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 2nd edition, Prentice-Hall, 2009.

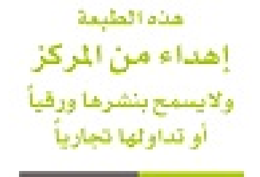
Kennedy, Graeme: *An Introduction to Corpus Linguistics*. London, Longman, 1998.

Kilgarriff, Adam, Rychly, Pavel, Smrz, Pavel, & Tugwell, David: The Sketch Engine. In the proceedings of *the Euralex, 6-10 July 2004*. Lorient, France, 2004.

Lee, Dong Ju: *Corpora and the classroom: a computer-aided error analysis of Korean students' writing and the design and evaluation of data-driven learning materials*. Unpublished thesis, University of Essex, 2007.

Leech, Geoffrey: Teaching and Language Corpora: a Convergence. In A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (Eds.), *Teaching and language corpora*. London, Longman, 1997, pp. 1-23.

Ljung, Magnus: Swedish TEFL Meets Reality. In Johansson, S. and Stenstrom, A-B. (Eds.) *English Computer Corpora: Selected Papers and Research Guide*. Mouton de Gruyter, Berlin and New York, 1991, pp. 245-256.



Longman Corpus Network: *The Longman Learners› Corpus*. 2012, Retrieved from: http://www.pearsonlongman.com/dictionaries/corpus/learners.html.

Meurers, Detmar: Learner Corpora and Natural Language Processing, *Cambridge Handbook of Learner Corpus Research*. edited by Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier. Cambridge University Press, Forthcoming, Retrieved from: http://www.sfs.uni-tuebingen.de/~dm/papers/Meurers-LCNLP-draft.pdf

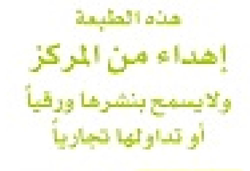
Milton, John, and Nandini, Chowdhury: Tagging the interlanguage of Chinese learners of English. In L. Flowerdew and A. K. K. Tong (Eds.), *Entering Text*. Hong Kong, The Hong Kong University of Science and Technology, 1994.

Nesselhauf, Nadja: Learner Corpora and Their Potential in Language Teaching. In: Sinclair, J. (Ed.), *How to Use Corpora in Language Teaching*. Amsterdam & Philadelphia, Benjamins, 2004, pp. 125-152.

Norvig, Peter: *How to Write a Spelling Corrector.* 2007, Retrieved from: http://norvig.com/spell-correct.html.

Packard, V. : Producing a concordance-based self-access vocabulary package: some problems and solutions. In: Flowerdew, Lynne & Tong, Anthony. (Eds.), *Entering Text*. Language Centre, University of Science and Technology, 1994, pp. 215-226.

Pravec, N. A. : Survey of learner corpora. *ICAME Journal*, 26, 2002, pp. 81-114, Retrieved from: http://icame.uib.no/ij26/pravec.pdf.



Scott, Mike: *WordSmith Tools version 6*. Liverpool: Lexical Analysis Software, 2012, http://www.lexically.net/wordsmith.

Simpson, R. C., Briggs, S. L., Ovens, J. and Swales, J. M. : *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI, The Regents of the University of Michigan, 2002, Retrieved from the University of Michigan, English Language Institute Web site: http://micase.elicorpora.info/about-micase

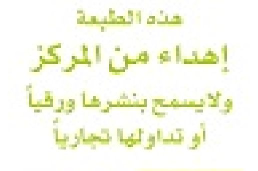
Sinclair, John: Corpus and Text - Basic Principles. In M. Wynne (Ed.), Developing Linguistic Corpora: a Guide to Good Practice. Oxford, Oxbow Books, 2005, pp. 1-16.

Sinclair, John: *Corpus, Concordance, Collocation*. Oxford, Oxford University Press, 1991.

Sinclair, John: *EAGLES. Preliminary recommendations on Corpus Typology,* 1996, Retrieved from: http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html

Sosnina, Ekaterina: *Russian Learner Translator Corpus (RusLTC),* 2014, Retrieved from: http://rus-ltc.org

Spence, Robert: A corpus of student L1–L2 translations. In S. Granger & J. Hung (Eds.), *Proceedings of the International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Hong-Kong, The Chinese University of Hong Kong, 1998, pp. 110–112.



Summers, Della: *Longman Essential Activator.* Longman. England, 1997.

Thoday, Elizabeth: Issues in building learner corpora: an investigation into the acquisition of German passive constructions. In *the 2nd Newcastle Postgraduate Conference in Theoretical and Applied Linguistics (25 June 2007)*. Newcastle, UK, 2007.

Tribble, Chris and Jones, Glyn: *Concordances in the Classroom: A Resource Book for Teachers*. Houston, TX, Athelstan, 1997.

Turton, Nigel D. and Heaton, J.B. : *Longman dictionary of common errors,* Pearson Longman. England, 2nd edition, 1996.

Zaghouani, Wajdi, Mohit, Behrang, Habash, Nizar, Obeid, Ossama, Tomeh, Nadi and Oflazer, Kemal: (Large-scale Arabic Error Annotation: Guidelines and Framework. In proceedings of the International Conference on *Language Resources and Evaluation (LREC 26-31 May 2014). Rejkavik, Iceland,* 2014.

المبحث الثالث

تصميم المدونات اللغوية وبناؤها

عبدالمحسن بن عبيد الثبيتي
مدينة الملك عبدالعزيز للعلوم والتقنية
الرياض، المملكة العربية السعودية
aalthubaity@kacst.edu.sa

مستخلص

بغض النظر عن التعريفات المتعددة للمدونات اللغوية (المدونة فيما بعد) باختلاف توجهات الباحثين في هذا المجال، فإنها بيانات يحاول الباحثون بواسطتها دراسة اللغة من خلال استخدامها الفعلي كما فعل علماء اللغة من قبل، وما زالوا في دراستهم للغة ووضعهم لقواعدها ولكن بمنهج جديد يعتمد على شواهد كثيرة أصبحت متاحة هذا العصر بسبب تطور الحاسب وتوافر النصوص الإلكترونية بشكل كبير. ولأنه من المستحيل عمليا جمع كل ما قيل وكُتب للحصول على أحكام دقيقة وفاصلة، فإننا أمام خيار آخر يبدو أكثر واقعية ويستند إلى أساس علمي ألا وهو جمع عينة متوازنة وممثلة للغة أو إحدى صورها أو ظواهرها محل الدراسة. ومثلما هو حاصل في مجالات علمية مختلفة، مثل: الطب أو علم النفس؛ فإن النتائج المتحصل عليها اعتمادا على دراسة العينات وتحليلها يمكن الوثوق بها متى ما توفرت ثلاثة شروط رئيسة، هي: ١) ألا تكون العينة منحازة وأن تكون كافية للدراسة، ٢) أن تدرس هذه العينة وتحلل باتباع منهج علمي، ٣) الحصول على نفس النتائج متى ما استخدم المنهج نفسه على ذات العينة. يقدم هذا الفصل إطارا عمليا للإجراءات التي تمكّن الباحث من تحقيق الشرط الأول من خلال اتباع معايير واضحة لتصميم وجمع نصوص المدونة لتكون عينة متوازنة وممثلة لمجال الدراسة اللغوية مع تعريف مبسط لبعض الأدوات التي تساعد في دراسة المدونات بشكل ميسر متى ما توفر المنهج العلمي لدى الباحث (الشرط الثاني). ويمكننا تحقيق الشرط الثالث عندما تكون جميع المعلومات الخاصة بتصميم وبناء المدونة موثقة ومتاحة للباحثين الآخرين مع إتاحة نصوص المدونة مجاناً أو بمقابل مادي- بما لا يخرق قوانين وأنظمة الحقوق الفكرية أو الخصوصية الشخصية - ليتمكن الباحثون من التحقق من نتائج الدراسات الأخرى التي أجريت على المدونة.

مقدمة

قبل أن يفكر الباحث في تصميم وبناء مدونته اللغوية الخاصة، عليه أن يجيب عن السؤالين التاليين:

١. ما الغرض الذي من أجله أريد أن استخدم المدونة؟

٢. هل توجد مدونة أو مدونات تجيب عن أسئلة بحثي؟

عادة ما يتم تحديد الغرض من استخدام المدونة بوضع سؤال أو عدة أسئلة تكون مدار البحث في المدونة، وتوجهُ هذه الأسئلة الباحث لنوعية المعلومات التي يريدها وكيفية الحصول عليها. وتتطلب الإجابة عن السؤال الثاني معرفةً بتصميم ومحتويات المدونات المتوفرة وأدوات البحث فيها واطلاعاً على الأبحاث التي تمت باستخدام هذه المدونات؛ ليطلع الباحث على جوانب القوة والقصور فيها، وليعرف مدى قدرتها على الإجابة عن أسئلة بحثه. وإن لم يجد الباحث مدونة تفي بالإجابة عن أسئلته فقد يضطر إلى أن يعدّل أسئلته أو محددات بحثه ليتمكن من استخدام المدونات المتوفرة، أو أن يغيرها تماما لتتناسب مع ما هو متوفر، أو قد يلجأ - إن كان لديه الوقت والإمكانات - إلى بناء مدونته الخاصة.

وعلى الرغم من توافر عدد لا بأس به من المدونات العربية (انظر المبحث الأول من هذا الكتاب) إلا أن أغلبها لا يوضح المعايير التي اُتبعت في تصميمها، وجمع نصوصها، وما هو الفرق بين التصميم والمحتوى الفعلي للمدونة بشكل واضح، ولعلي أستثني من هذا ثلاثاً من المدونات العربية وهي مدونة اللغة العربية المعاصرة Corpus of Contemporary Arabic من جامعة ليدز (السليطي وأتول ٢٠٠٦، Al-Sulaiti and Atwell)، والمدونة العالمية للغة العربية International Corpus of Arabic من مكتبة الإسكندرية (الأنصاري وآخرون

Alansari et al ،٢٠٠٧.)، والمدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية (الثبيتي ٢٠١٤-Al-Thubaity). وأفرّق هنا بين أن توضع معايير تستخدم مرشدًا لجمع نصوص المدونة واستخدامها لجمع النصوص وبين أن يتم توضيح محتويات المدونة وتوزيعها بعد جمع نصوصها دون الاسترشاد بأية معايير تضبط هذا الجمع.

يتطرق القسم الثاني من هذا المبحث بشكل مبسط ومختصر إلى أهم المعايير التي من المفترض وضعها وتوثيقها عند تصميم المدونات لتكون متوازنة وممثلة للغة محل الدراسة. ويتعرض القسم الثالث إلى خطوات بناء المدونات بحيث تحقق معايير التصميم. أما القسم الرابع فيتعرض لأهم الأدوات اللازمة لمعالجة المدونات. وفي القسم الخامس أضع مثالاً تطبيقيا لما ذكر في القسمين الثاني والثالث. وفي القسم السادس مثال تطبيقي على خطوات التصميم والبناء المذكورة في القسمين الثالث والرابع. أما خاتمة المبحث وخلاصته فهي في القسم السابع.

معايير التصميم

يتطرق هذا القسم إلى أهم المعايير التي يجب وضعها في الاعتبار وتحديدها بوضوح قبل البدء بجمع نصوص المدونة. تساعد هذه المعايير -إن طبقت بشكل دقيق عند جمع النصوص- على أن تكون المدونة قادرة على إجابة أسئلة البحث أو الغرض الذي بُنيت من أجله. تُعد هذه المعايير إطارًا عامًّا يحدد نوعية وكمية الجهد الذي يجب أن يبذل في الخطوات التي تلي التصميم. إنّ وضع هذه المعايير والاهتمام بها منذ البداية ينقل المدونات من كونها بيانات يهتم بها بعد إنجازها «مجموعة من البيانات اللغوية المكتوبة أو المنطوقة» كما عرفها كريستال (كريستال ١٩٩٢، Crystal) إلى كونها عملاً منهجياً مخططاً ومدروساً له أهداف واضحة منذ البداية «مجموعة من نصوص اللغة في صورة

إلكترونية تجمع اعتمادًا على معايير خارجية؛ لتمثل قدر المستطاع اللغة أو أحد صورها لتكون مصدرًا للأبحاث اللغوية» (سنكلير ٢٠٠٥، Sinclair).

والمقصود بالمعايير الخارجية هنا، المعايير التي تعتمد على الوظيفة التواصلية للنص في المجتمع الذي ظهر فيه، وهذه المعايير تشبه إلى حد كبير المعلومات الببليوغرافية للنص مثل الوعاء الذي صدر فيه النص وموضوعاته المختلفة والمنطقة الجغرافية التي صدر منها، وقد تشمل معلومات أخرى مثل جنس الكاتب وجنسيته ومستواه الاجتماعي. أما المعايير الداخلية للنص فإنها تتعلق بالبنية اللغوية الداخلية للنص ذاته مثل كونه يحتوي على تركيب نحوي أو صرفي بصيغة معينة. إن الاعتماد على المعايير الداخلية يعزز بشكل كبير ظهور هذه التراكيب وما يلازمها عادة وبالتالي لا يعكس صورة محايدة عن استخدامها الفعلي مقارنة بغيرها من الظواهر اللغوية.

إن أول ما يتحكم في معايير بناء المدونات التي سوف نتطرق إليها في الأقسام الفرعية التالية هو الغرض الذي تُبنى من أجله، هذا الغرض يجب أن يكون واضحاً ومحدداً بدقة منذ البداية. ويمكننا بشكل عام أن نقسم الأغراض التي تبنى لأجلها المدونات إلى قسمين رئيسين: أغراض خاصة/محددة وأغراض عامة/شاملة.

يتميز القسم الأول بمحاولته الإجابة عن أسئلة محددة، وبالتالي فإن نتائج الدراسة لا يمكن تعميمها على اللغة ككل، مثل: أن يكون الغرض من بناء المدونة دراسة التراكيب النحوية في لغة الشعر الجاهلي أو دراسة الأفعال في أبحاث علوم الكيمياء أو دراسة أخطاء متعلمي اللغة العربية لغة ثانية، ويمكن أن يلحق بهذا القسم المدونات التي تبنى لأغراض معالجة اللغة أو نمذجتها فقط ولم يوضع في الحسبان عند تصميمها الدراسات اللغوية. ومن أمثلة هذه المدونات مدونة متعلمي العربية من جامعة ليدز (الفيفي وآخرون ٢٠١٤، Alfaifi et al.) (انظر المبحث الثاني من هذا الكتاب)، ومدونة مدينة الملك عبدالعزيز للعلوم

والتقنية لتصنيف النصوص (خورشيد والثبيتي ٢٠١٣، Khorsheed and Al-Thubaity)، ومدونة تصحيح الأخطاء الإملائية (الكنهل وآخرون ٢٠١٢، Alkanhal et al.).

أما القسم الثاني من الأغراض التي تبنى لأجلها المدونات فيتميز بشكل عام بتنوع موضوعاته ومحاولته الوصول إلى مدونة تستطيع الإجابة عن أسئلة متنوعة، ويندرج تحت هذا النوع من المدونات المدوناتُ المرجعية. ومن أمثلة هذه المدونات المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية، ومدونة آر تن تن arTenTen على موقع سكتش إنجن(٧) Sketch Engine.

فيما يلي شرح لأهم معايير تصميم المدونات التي جمع شتاتها سنكلير (سنكلير ٢٠٠٥) وكانت نتاجاً لخبرته الطويلة في العمل على المدونات وإشرافه المباشر على المدونة الشهيرة بنك اللغة الإنجليزية The Bank of English مع بعض النقاش والرجوع إلى مراجع أخرى يتم الإشارة إليها في حينها حسبما تقضي الحاجة. وهذه المعايير هي:

٣. ١ لغة المدونة

لا يقتصر تحديد لغة المدونة على التحديد العام للغة مثل كونها اللغة العربية أو الإنجليزية، بل يتعدى ذلك إلى تحديد تفاصيل أكثر دقة. ومن أمثلة هذه التفاصيل أن تكون لغة المدونة هي اللغة الفصحى المعاصرة أو القديمة التراثية (إن جازت هذه التسمية) أو حتى اللهجات إن كان هذا هو مطلب الدراسة، كما أن التفاوت داخل هذه الأنواع الثلاثة يجب أن يُنظر إليه ويؤخذ بالحسبان أيضاً. فاللغة العربية المعاصرة -على سبيل المثال- على الرغم من أن لها ملامح عامة عند متحدثيها في البلدان العربية إلا أن هناك اختلافات واضحة على المستوى اللفظي -على الأقل- فيما بينهم. ومن أمثلة هذه الاختلافات اللفظية في العربية المعاصرة بين لغة الصحافة السعودية والمغربية كلمتا «بنوك» و «أبناك».

٣. ٢ طبيعة النصوص

تختلف الطبيعة التي تظهر فيها اللغة البشرية، فهي قد تكون منطوقة وهذا الأغلب وقد تكون مكتوبة في صور متعددة وقد تكون أيضًا لغة إشارة، ومهما يكن الأصل الذي ظهرت فيه اللغة فيجب تحويل هذا الأصل إلى صورة إلكترونية قابلة للمعالجة الآلية، وفي الأغلب فإن هذه الصورة تكون على صيغة TXT.

إن تحديد نسبة ما سوف تحتويه المدونة من نصوص منطوقة أو مكتوبة أو كليهما له أثر كبير على اختيار الأوعية والفترات الزمنية وكذلك في الجهد الذي سوف يستغرق في جمع محتويات المدونة، فالجزء المنطوق من المدونة يستغرق وقتاً أطول بكثير في الحصول عليه ومن ثم تحويله إلى نص مكتوب مقارنة بالجزء المكتوب أصلاً، وتزداد الصعوبة لو كانت اللغة هي اللغة الفصحى، ناهيك عن التكلفة المادية الباهظة لذلك والاحترازات اللازمة لعدم خرق الخصوصية الشخصية.

٣. ٣ تاريخ النصوص

تختلف الفترات الزمنية التي يجب أن تغطيها المدونة باختلاف الأغراض التي تبنى من أجلها، فالمدونات التي تسعى لدراسة اللغة الحديثة أو المتخصصة عادة ما تتميز بقصر الفترات التي تغطيها، وتتراوح هذه الفترات من سنة إلى عدة سنوات. وعلى النقيض من ذلك المدونات التي تسعى لدراسة تطور اللغة إذ تتضمن نصوصاً من فترات زمنية طويلة تتراوح بين عشرات السنين إلى المئات منها. وبالطبع فإن طول الفترة الزمنية التي تغطيها المدونة يميزها عن غيرها من المدونات التي تغطي فترات أقصر وخصوصا فيما يتعلق بالمدونات المرجعية Reference Corpora سواء كانت حديثة أو تاريخية. ويجب مراعاة أن تتضمن المدونة نصوصا تغطي جميع أجزاء الفترة الزمنية لا أجزاء متفرقة منها قدر الإمكان، وتسعى بعض المدونات التي تهتم بأغراض مقارنة اللغة بين

فترة زمنية وأخرى إلى تضمين نصوص من فترات زمنية مختلفة ومتباعدة مثل أن تحوي المدونة نصوصًا من السنوات العشر الأولى من القرنين العشرين والواحد والعشرين الميلاديين.

٣. ٤ المنطقة الجغرافية

يقصد بالمنطقة الجغرافية هنا البلد أو البلدان التي صدرت فيها النصوص، وقد تستدعي الحاجة تحديد المناطق المختلفة داخل البلد الواحد نفسه، فعلى سبيل المثال لو كان الغرض إنشاء مدونة لغوية تختص بالمملكة العربية السعودية فإنه من المفترض الأخذ بعين الاعتبار جميع مناطق المملكة.

وتزداد قيمة المدونة عند تضمينها نصوصاً من بلدان ومناطق مختلفة تتحدث نفس اللغة بحيث تكشف الاختلافات اللغوية والثقافية بين هذه البلدان أو المناطق، فعلى سبيل المثال لا يمكن لمدونة أن تكشف عن الأنماط اللغوية المشتركة أو المختلفة في اللغة العربية مالم تُضمن نصوصا من جميع البلدان العربية، ولكن التنوع في البلدان أو المناطق ليس شرطًا واجبًا بل إن الحاجة لهذا التنوع يمليه الغرض من المدونة فحسب.

٣. ٥ الوعاء

تظهر النصوص في أوعية مختلفة مثل الصحف والمجلات والكتب والرسائل الجامعية أو الشابكة، ولكل من هذه الأوعية سماته اللغوية العامة وخصائصه التي يمكن أن تميزه عن غيره على المستوى اللفظي والتركيبي، فلغة العلم في الرسائل الجامعية والدوريات المحكمة تتميز بالدقة والوضوح واستخدام المصطلحات العلمية، ولا تستخدم المجاز، فلا يمكن التعبير عن فكرة أو معنى معين إلا بمصطلح واحد وثابت متفق عليه في الأغلب؛ بينما الحال مختلف في لغة الصحافة والأدب مثلاً.

إن تحديد أوعية متنوعة لتُضمّن في المدونة يزيد من قيمتها وفائدتها لدراسات وتطبيقات مختلفة بخلاف المدونات المنحصرة في وعاء واحد، وعلى العموم فإن تعدد الأوعية يُعد من علامات المدونات المرجعية في الأغلب.

٣. ٦ المجال

لكل وعاء من الأوعية مجالات تختص به وقد تظهر فيه فقط ولا تظهر في غيره، فعلى سبيل المثال نجد الأخبار والمقالات والتقارير بوصفها مجالات مختلفة يمكن أن تكون في الصحف والمجلات ولا يمكن أن توجد في الكتب أو الرسائل الجامعية. وتتنوع اللغة المستخدمة داخل الوعاء الواحد باختلاف مجالاته، فالدوريات العلمية المحكمة على سبيل المثال تعد وعاء يجمع عدة مجالات مختلفة مثل أصول الفقه والطب والهندسة، وعلى الرغم من أن هذه المجالات قد تتسم بطابع واحد عام وهو استخدام لغة العلم إلا أنها تتفاوت في طرق التعبير والمصطلحات المستخدمة. إن التباين في الأفكار بين هذه المجالات والطرق المستخدمة في التعبير عنها يثري المدونة ويسمح بظهور الأنماط العامة للغة العلم وكذلك الأنماط الخاصة بكل مجال، كما أن معرفة هذه المجالات وتحديد نسبتها من كل وعاء يساعد في التخطيط المبكر لجمع نصوص المدونة وتيسيره فيما بعد.

٣. ٧ حجم العينة

في هذا المعيار يتم تحديد ما إذا كانت المدونة سوف تتضمن نصوصًا كاملة أم أجزاء من النصوص، ويوجد رأيان مختلفان بهذا الخصوص: فالرأي الأول يرى تضمين النصوص كاملة ما أمكن (سنكلير ٢٠٠٥)؛ لأن وجود النص كاملا يزيد من فرصة ظهور خواص لغوية متعددة ومختلفة ومترابطة في سياقها الطبيعي، وفي المقابل يرى الرأي الثاني أن اختيار جزء من النص يزيد من تنوع مواد المدونة، ناهيك عن العوائق المتعلقة بحقوق الملكية الفكرية وصعوبة

الحصول على النصوص كاملة (ماير ٢٠٠٢). وعلى الرغم من ذلك فيجب عدم الاكتفاء بجزء واحد من النص، بل بأجزاء متعددة من أوله وأوسطه وآخره (ماكنري وآخرون ٢٠٠٦). وبطبيعة الحال فإن مثل هذا القرار له علاقة بالنصوص ذات الحجم الكبير مثل الكتب والرسائل الجامعية وليس النصوص ذات الحجم الصغير مثل نصوص الصحف.

ومهما يكن فإن تحديد حجم عينة النصوص مرتبط بحجم المدونة، فقد يكون من المناسب أن تكون عينة النصوص أجزاء من النص عندما تكون المدونة صغيرة الحجم ومن الممكن الوصول للحجم بهذه الطريقة، أما إن كانت المدونة كبيرة الحجم فالأولى تضمين النص كاملاً، مع التأكيد على وجوب الحرص على استخدام نفس حجم العينة في جميع النصوص ما أمكن.

٣. ٨ حجم المدونة

يقاس حجم المدونة بعدد كلماتها. والكلمة هنا تعني أي مجموعة متتابعة من الرموز لا يفصل بينها فراغ، وبالتالي فإن بعض الكلمات -حسب هذا التعريف- قد تكون كلمة معروفة وصحيحة مثل «كتاب»، أو قد تكون أرقاما «٩٧٣٠» أو كلمات ليس لها معنى «ععخعج» أو كلمات تحوي أخطاء طباعية «كعنبوت» أو كلمات تمت إضافة الكشيدة(٨) في وسطها «بســــم». وبالتالي فإن «بســــم» و«بسم» تعدان كلمتين مختلفتين بالنسبة لأدوات معالجة المدونات لاختلاف شكليهما على الرغم من كونهما كلمة واحدة. مثل هذه الأمثلة موجودة في أغلب المدونات خصوصاً الكبيرة منها، ونسبتها إلى إجمالي حجم المدونة لا يذكر ولكن الإشارة إليها لازمة لفهم معنى الكلمة التي بناء عليها يتم قياس حجم المدونة.

ولا بد من التنبيه على أن الآراء تتفاوت حول حجم المدونة، فهناك من يرى أنه كلما ازداد حجم المدونة كان ذلك أفضل (سنكلير ١٩٩١، Sinclair)،

والسبب في ذلك أن أغلب كلمات المدونة لا تتكرر بالشكل الكافي، ولذا فإنه كلما زاد الحجم، سنحت الفرصة لظهور أنماط أكثر ثباتاً وقبولاً. ومع ذلك يرى عدد من الباحثين أن مدونة بحجم مليون كلمة كافية للأبحاث اللغوية التي تبحث عن الظواهر العامة في اللغة (انظر مناقشة هذا الموضوع في السليطي وأتول، ٢٠٠٦).

كما أن حجم المدونة يتعلق بنوعية العمل أو الدراسات المطلوبة، فالمدونات المتعلقة بالدراسات المعجمية تتطلب حجما أكبر مقارنة بالمدونات المتعلقة بالتراكيب النحوية، وقد ساعد التطور الحاصل في تقنية المعلومات والتوفر المتزايد للنصوص الإلكترونية على بناء مدونات لغوية كبيرة بصورة أسهل من ذي قبل.

وبشكل عام، فإن التحديد الدقيق للحجم الكافي لتحقيق الغرض الذي من أجله تبنى المدونة أمرٌ صعب للغاية ولا توجد معادلة رياضية لذلك؛ بل يخضع بشكل كبير للخبرات السابقة لمصمم المدونة بالإضافة إلى الخبرات المنشورة في هذا المجال مما هو مقارب لغرض المدونة ويتضح بصورة أكبر بعد استخدام المدونة.

ويمكن تحديد نسبة أو عدد الكلمات التي تمثل كل معيار من معايير تصميم المدونة بطريقتين: الطريقة الأولى تعتمد على التوزيع من الأعلى إلى الأسفل، حيث يتم تحديد الحجم الكلي للمدونة اللغوية ثم تحديد نسبة الجزء المكتوب والمنطوق منها، يتبع ذلك تحديد النسبة المناسبة لكل وعاء من أوعيتها، ثم يتم تقسيم هذه النسبة على مجالات الوعاء ثم موضوعاته، يتبع ذلك تحديد نسبة كل فترة زمنية ومنطقة جغرافية إن كانت المدونة تغطي عدة بلدان وفترات زمنية مختلفة، ويمكن بدلاً من ذلك توزيع الحجم الكلي للمدونة على الفترات الزمنية ثم البلدان ثم الأوعية والمجالات والموضوعات.

أما الطريقة الأخرى فتعتمد على التوزيع من الأسفل إلى الأعلى، فيتم فيها تحديد عدد الكلمات المناسب لكل موضوع ونسبة كل بلد وفترة زمنية من هذا الموضوع، وبالتالي فإن مجموع عدد كلمات كل موضوع يعطي العدد الإجمالي للمجال، ويكون مجموع عدد كلمات كل وعاء هو مجموع عدد كلمات كل مجالاته، وبذلك يمكن تحديد الحجم الكلي للمدونة اللغوية بجمع عدد كلمات كل وعاء فيها.

٣. ٩ معايير أخرى

إن المعايير السابقة تُعد أهم المعايير التي يجب وضعها في الحسبان عند تصميم المدونة، وهناك معايير أخرى يجب اعتبارها في بعض المدونات حسب الغرض الذي من أجله يتم بنائها. من هذه المعايير تحديد جنس الكاتب ذكراً أو أنثى، وكم النسبة التي يجب أن تحتويها المدونة من نتاج كلٍ من الجنسين. ومثل هذا المعيار مهم عندما يكون غرض المدونة الدراسات المقارنة بين كتابات الرجال والنساء. ومن الأمثلة الأخرى تحديد الفئة العمرية والمستوى التعليمي والاجتماعي أو حتى عدد المتابعين للمغرد إن كانت المدونة تدرس لغة الخطاب في تويتر Twitter. وهذه المعايير التي تم ذكرها إنما هي أمثلة فقط لبعض المعايير الخاصة التي يتطلبها تصميم مدونات ذات مواصفات وأغراض خاصة ولا تعني الحصر.

٣. ١٠ التمثيل والتوازن

المقصود بالتمثيل هو قدرة المدونة على تمثيل اللغة أو صورها المختلفة محل الدراسة، ويرى ماكنري وآخرون (٢٠٠٦) أن ما يميز المدونات عن أي مجموعة عشوائية من النصوص هو قدرتها على تمثيل اللغة ومتغيراتها، ويرتبط التمثيل ارتباطاً وثيقاً بحجم المدونة والتنوعات المختلفة فيها ونسبة كل تنوع، فكلما زاد

الحجم والتنوع كانت قدرة المدونة أكبر على تمثيل اللغة والإجابة عن أسئلة الدراسة.

أما المقصود بالتوازن فهو أن تمثل نصوص المدونة الواقع اللغوي كما هو في خارجها فلا تكون منحازة لمحتوى أو مستوى دون آخر، وليس بالضرورة أن يعني ذلك التساوي في حجم ما تحويه المدونة من كل وعاء أو مجال إلا إن كان ذلك لازماً مثلما هو الحال في الدراسات المقارنة. فلو نظرنا للنتاج الكتابي خارج المدونات فسنجد أن الغلبة للصحف مقارنة بالرسائل الجامعية، وعلى هذا يجب أن يكون حجم الصحف من أوعية المدونة المرجعية أكبر من الرسائل الجامعية بحسب الواقع فعلاً. وكذلك الحال عند التوزيع الجغرافي لمواد المدونة، فالدول التي تتميز بنتاج أكبر يجب أن يكون لها النصيب الأكبر من الحجم بحسب الواقع. وعلى الرغم من أن هذا الكلام صحيح في مجمله لكننا لا نستطيع أن نحدد بدقة تامة مقدار النسبة لكل جزء من محتوى المدونة.

إن التمثيل والتوازن مفهومان مهمان ومرتبطان لا يمكن فصلهما عن بعضهما، ويتحققان بشكل كبير عند مراعاة جميع المعايير السابقة بحسب الغرض من المدونة. وقد حظي هذان المفهومان باهتمام بالغ -ولا زالا- في كثير من الدراسات التي تناقشهما وتعرض العديد من وجهات النظر حولهما، انظر على سبيل المثال (سنكلير ١٩٩١) (أتكنز وآخرون ١٩٩٢، Atkins et al.) (بايبر ١٩٩٣، Biber) (بايبر وآخرون ١٩٩٨، Biber et al.) (ليتش ٢٠٠٦، Leech). وبسبب الاختلاف الكبير حول سبل تحقيق التمثيل والتوازن كانت دقة تمثيل المدونات الكبرى الشهيرة للغة الإنجليزية، كالمدونة الوطنية البريطانية British National Corpus محل تساؤل عند بعض الباحثين (أحمد ٢٠٠٨، Ahmad). وللتغلب على هذا الإشكال وصعوبة تحقيقه فعليا بشكل كامل اقترح تيوبرت وكيرماكوفا (تيوبرت و كيرماكوفا ٢٠٠٧، Teubert and Cermáková) استخدام المدونات السانحة Opportunistic Corpora؛ إذ تمتاز المدونات

السانحة بميزتين رئيستين، أولاهما: أنها كبيرة الحجم يجمع فيها كل ما يمكن جمعه من نصوص مختلفة، وثانيهما: أنها توفر معلومات تفصيلية شاملة عن النصوص التي تحتويها، وعندها يكون بإمكان الباحث اختيار النصوص التي تمثل من وجهة نظره التمثيل والتوازن لدراسته التي يرغب القيام بها.

ومهما تكن وجهة النظر حيال التمثيل أو التوازن فإنني أرى أنه لا توجد مدونة لغوية ممثلة للغة وصورها المختلفة بشكل كامل، وأن ما يمكن الحكم عليه بشكل أدق هو قدرة المدونة على تمثيل موضوع الدراسة أفضل من غيرها من المدونات الأخرى.

٣. ١١ البيانات الأساسية للنصوص

البيانات الأساسية للنصوص شبيهة إلى حد كبير بالمعلومات الببليوغرافية. وتحديد هذه المعلومات الأساسية وتوفيرها للباحثين مع نصوص المدونة يساعد في إدارة المدونة، وتحديد مدى مناسبتها لغرض الدراسة بشكل كبير، وكذلك يمكن الاستفادة منها في تحديد الأجزاء الأكثر مناسبة للدراسات اللاحقة، والاكتفاء بها، كما تساعد البيانات الأساسية للنصوص في الاستفادة من المدونة في المستقبل في بناء مدونات لغوية أخرى تتضمن هذه المدونة بكاملها أو أجزاء منها.

تشمل البيانات الأساسية للنصوص المعلومات التالية وليست حصرًا عليها:

عنوان النص.

اسم المؤلف وجنسيته وجنسه.

وعاء النص ومجاله وموضوعه.

تاريخ صدور النص.

ناشر النص والبلد الذي ينتمي إليه.

مصدر النص.

تاريخ إضافة النص للمدونة.

بناء المدونات

إن بناء المدونات لا يعني جمع النصوص حسب المعايير فحسب، بل يتطلب الأخذ بالاعتبار إجراءات عدة تسبق الجمع وتليه، وفيما يلي سوف نوضح هذه الخطوات مع إعطاء نماذج لما يمكن أن تتضمنه هذه الخطوات:

٣. أ حقوق الملكية الفكرية

إن حقوق الملكية الفكرية لنصوص المدونة هي أول ما يتوجب النظر إليه بحرص قبل الشروع في بناء المدونة؛ إذ إن انتهاك حقوق الملكية الفكرية لمالكي النصوص -جهاتٍ أو أفرادا- قد يعرّض الشخص أو الجهة المسؤولة عن بناء المدونة للمساءلة القانونية، وقد يجهض المشروع برمته خصوصاً عند توزيع نصوص المدونة سواء بمقابل أو بالمجان. وعلى الرغم من أهمية هذا الموضوع وحساسيته فإن التغلب على صعوباته ممكن في أغلب الأحيان.

قبل التفكير في الحصول على إذن مسبق لتضمين النصوص المحمية بقوانين الملكية الفكرية في المدونة من المستحسن البحث عن نصوص ليست محمية بهذه القوانين، ويتضمن هذا النوع النصوص التاريخية القديمة، والنصوص التي انتهت مدة حمايتها النظامية بموجب القانون، والنصوص التي ينص أصحابها على مجانية الحصول عليها وتوزيعها، وكذلك ما توفره بعض الجهات الرسمية من إصدارات ونشرات تعريفية وتثقيفية.

وتنص بعض قوانين حماية الملكية الفكرية على عدم خضوع بعض النصوص للحماية. فعلى سبيل المثال فإن نظام حماية حقوق المؤلف ولائحته التنفيذية(٩) في المملكة العربية السعودية ينص على استثناء بعض المصنفات من الحماية،

كالأنظمة والأحكام القضائية والوثائق الرسمية وما تنشره الصحف والمجلات من الأخبار اليومية أو الحوادث ذات الصبغة الإخبارية، كما تسمح بعض الأنظمة بالاستفادة من النصوص كاملة أو أجزاء منها لأغراض تعليمية أو بحثية. ومع هذا السماح بالاستفادة من النصوص إلا أنها لا تهمل وجوب الإشارة إلى اسم صاحب الحق في ملكية هذا النص. وقد لجأت بعض المدونات كبيرة الحجم إلى الاستفادة من هذا النظام، فلا تسمح بتوزيع نصوص المدونة ولا تعرض نصوصها للاطلاع بل تعرض السياقات التي تظهر فيه الكلمة فقط مع الإشارة الى مؤلف النص وكافة المعلومات المتوفرة عن هذا النص في المدونة.

وعندما تدعو الحاجة للحصول على إذن بتضمين نصوص محمية بموجب النظام، كحقوق الملكية الفكرية أو نصوص قد تنتهك الخصوصية الشخصية، كحوارات المرضى مع أطبائهم، يجب أن يفرق القائمون على المدونة بين موضوعين رئيسين، والحصول على الإذن بهما، وهما حق تضمين النصوص في المدونة، وحق توزيع أو نشر هذه النصوص مجاناً أو بمقابل مادي.

٣. ب تحديد المصادر

إن تحديد المصادر التي يمكن أن تجمع منها نصوص المدونة يوفر الكثير من الوقت والجهد عند القيام بالخطوة التالية وهي جمع النصوص، ويمكننا تقسيم هذه المصادر إلى قسمين رئيسين، الأول: المصادر التي توفر نصوصاً إلكترونية للجمع وللمعالجة المباشرة، مثل: مواقع الصحف والمجلات والمكتبات الإلكترونية التي توفر نصوصا بصيغة نصية (DOC, DOCX, TXT). والآخر: المصادر التي توفر نصوصا قابلة للجمع ولكنها غير قابلة للمعالجة الإلكترونية مباشرة، بل تتطلب جهدا إضافيا لتحويلها لصيغة نصية، كأن تكون نصوصها على صيغة صور أو PDF أو أن تكون ورقية، ويمكن في هذه الحالة استخدام برامج التعرف الضوئي على الحروف لتحويلها إلى صيغة نصية إلكترونية أو كتابة النص وحفظه بصيغة نصية.

ويتطلب تحديد المصادر إعداد قائمة بها وبروابطها الإلكترونية على الشابكة أو طريقة الحصول عليها إن كانت ورقية، مع تسجيل أي ملاحظات تخص رخص الحصول عليها، وتضمينها في المدونة. وتعتبر هذه القائمة، قائمة أولية يجب اختبارها وتجريبها وجمع نصوص متعددة منها؛ لمعرفة الأفضل والأسهل في جمع النصوص، كما يجب الإضافة إليها وتنويعها بقدر الإمكان وحذف ما لا يصلح منها لأي سبب، كعدم توفر معلومات دقيقة عن النصوص مثل نسبة نص لغير صاحبه.

٣. ج الجمع

اعتماداً على قائمة المصادر التي تم تحديدها مسبقا يبدأ العمل في جمع نصوص المدونة. وتوفيراً للوقت والجهد والتكلفة يجب التركيز على القسم الأول من المصادر (المصادر التي توفر نصوصاً إلكترونية للجمع وللمعالجة المباشرة) وعدم العمل على القسم الثاني (المصادر التي توفر نصوصا قابلة للجمع ولكنها غير قابلة للمعالجة الإلكترونية المباشرة) إلا في أضيق الحالات. ومن الممكن في بعض الحالات أن يتم جمع النصوص آلياً حيث توجد عدة برامج يمكن أن تقوم بهذا العمل كبرنامج بوت كات (١٠)BootCat على سبيل المثال لا الحصر. وقد استُخدمت هذه الطريقة بالكامل لبناء بعض المدونات، انظر على سبيل المثال (الزهراني ٢٠١٣، Alzahrani) (خوجه ٢٠٠٩، Khoja) (جاكوبيكيك وآخرون ٢٠١٣، Jakubíček et al.). ومع ذلك، يجب التنبه إلى أن نتاج الجمع الآلي للنصوص بحاجة إلى مراجعة دقيقة؛ فبعض هذه النصوص قد يكون مليئاً بالأخطاء، وبعضها قد يكون مكرراً أو يحتوي على بيانات ليس لها علاقة بالنص الأصلي، كروابط لصفحات أخرى أو إعلانات تجارية أو بيانات لها علاقة بموقع الشابكة. وعلى الرغم من فائدة الطريقة الآلية في الجمع السريع لنصوص المدونة إلا أنه من الصعب تحديد الأوعية الخاصة بالنصوص ومجالاتها والتواريخ التي ظهرت فيها ومن هم مؤلفوها، لذا فإن تعيين المواقع

التي تجمع منها النصوص لتحديد الأوعية والموضوعات يجب أن يكون محل الاهتمام. ولضمان جودة نواتج عملية جمع النصوص يفضل أن تُجزأ عملية الجمع إلى مراحل يتم تقسيمها بناء على الأوعية مثلا، وكلما انتهى قسم تتم مراجعة نصوصه والتأكد من تحقيقه لمعايير التصميم.

٣. د الترميز والتسمية والحفظ

بعد جمع النصوص، ولزيادة الاستفادة من المدونة وتسهيل إدارتها من المستحسن القيام بالتالي:

توحيد ترميز النصوص، ومن المفضل تحويلها كاملة إلى ترميز واحد يكون مقبولاً من أغلب أنظمة التشغيل وبرامج معالجة المدونات، مثل:UTF8 و UTF16، فاستخدام ترميز الوندوز Windows مثلاً قد لا يساعد على الاستفادة من المدونة في حال معالجتها بأنظمة مختلفة عن الوندوز.

توحيد طريقة التسمية، ومن المفضل استخدام طريقة تعتمد على الأرقام والأحرف اللاتينية، فمثلاً يمكن تقسيم اسم الملف إلى عدة خانات يفصل بينها «-» حيث تعبر كل خانة منها عن أحد معايير التصنيف، فيكون لكل بلد ووعاء ومجال وموضوع وفترة زمنية رمزه الخاص وتكون آخر خانة للرقم التسلسلي للملف.

حفظ الملفات في مجلدات منفصلة حسب وعائها أو فترتها الزمنية أو البلد الذي صدرت فيه مع حفظ القائمة المشتملة على اسم النص ومعلوماته حسب معايير التصميم. وعندما تكون المدونة كبيرة الحجم وتحوي عدداً كبيراً من النصوص فقد يكون من الأفضل حفظ النصوص وبياناتها الأساسية في قاعدة بيانات، ليمكن عند ذلك إدارتها والتحكم فيها بسهولة، كما يمكن بهذه الطريقة تصديرها بصيغ متعددة حسب الرغبة متى ما دعت الحاجة إلى ذلك.

٣. هـ التحشية

من المستحسن -وقد يكون لازماً في بعض الحالات- أن يتم إثراء نصوص المدونة بمعلومات تزيد من فائدتها، وتساعد في إجراء المزيد من الدراسات المتعمقة. وتنقسم هذه المعلومات إلى ثلاثة أقسام:

القسم الأول: يتضمن معلومات عن النص نفسه وهي التي أطلقنا عليها من قبل مسمى البيانات الأساسية للنصوص، ويتم ذلك باتباع عدة منهجيات وأساليب معيارية تساعد في تبادل المدونات وسهولة استخدامها بين الباحثين. ومن أبرز هذه المعايير مبادرة ترميز النصوص(١١) Text Encoding Initiative (TEI)، والمرجع القياسي لترميز المدونات(١٢) Corpus Encoding Standard (CES). وهناك من يرى أن هذا القسم ليس له علاقة بالتحشية فيضع له مصطلحاً آخر وهو التعليم (من العلامة) Mark-up (ماكنري وآخرون، ٢٠٠٦).

القسم الثاني: معلومات تتعلق ببنية النص وتركيبه الظاهريين، مثل تحديد نهاية وبداية الفقرات والجمل والعبارات داخل النص، وهذه التحشية تساعد في الدراسات التي تنظر في العلاقات بين الجمل في الفقرات الواحدة وبين الفقرات في النص ذاته، كما تكون أساساً لبناء البنوك الشجرية Treebanks والتحليل النحوي Parsing. ويمكن أيضا كما في حالة الأوراق العلمية مثلاً تحديد بداية ونهاية الأجزاء المكونة للورقة العلمية مثل العنوان والكلمات المفتاحية والمقدمة والدراسات السابقة والتجارب ومناقشة النتائج والخاتمة، ومثل هذه التحشية تفيد في الدراسات التي تستهدف معرفة الخواص اللغوية لمثل هذا النوع من الكتابات وكيف تترابط أجزاءها.

القسم الثالث: ما يتعلق بإضافة نتائج التحليل اللغوي للنصوص كإضافة الوسوم النحوية والوسوم الدلالية وإضافة معلومات الإحالة بين الضمائر

والأسماء وكذلك معلومات التحليل النحوي. وعلاوة على ذلك يمكن للمدونات اللغوية العربية أن تشمل تحشيات من نوع آخر، كالتحشية الخاصة بالجذور وأنواعها والتوسيم الصرفي. ويجب التنويه إلى أن هذا النوع والأنواع الأخرى من التحشية يجب أن تتبع منهجا واحداً ودقيقاً في تحديد معلومات التحشية وذلك بوضع قائمة محددة سلفاً فلا يستخدم غيرها ولا تهمل إن وجدت.

وبطبيعة الحال، فإن عمليات التحشية في القسمين الثاني والثالث يمكن أن تتم بصورة آلية، وهذا هو المتبع في غالب المدونات الإنجليزية بسبب دقة مثل هذه الأنظمة الآلية في اللغة الإنجليزية؛ ولكن الحال في العربية مختلف، فالأبحاث في هذا المجال محدودة والموجود منها لا يراعي ما هو مستقر في النحو العربي إجمالاً (الثبيتي ٢٠١٤). وعلى الرغم من ذلك فالاستفادة من أنظمة التوسيم النحوي المتوفرة حاليا ممكنة وقد تكون ذات فائدة متى ما أدرك الباحث جوانب النقص فيها، وكان مطلعاً على قائمة الوسوم النحوية فيها، وعالماً بمدلولاتها، وكانت كافية ومحققة لغرضه البحثي.

الأدوات

مهما كانت الجهود والأوقات والأموال التي تصرف في تصميم المدونة وجمع نصوصها لتوائم التصميم فإنها ستكون بيانات بلا فائدة مالم يكن هناك أدوات قادرة على البحث فيها واستعراض نتائج هذا البحث والمساعدة في تحليله كماً ونوعاً للنظر أولاً في مدى تحقيق المدونة لغرض الدراسة، ولإجراء التحليلات اللغوية المختلفة على بياناتها ثانياً. وسوف نستعرض باختصار أهم الوظائف اللازمة في أنظمة معالجة المدونات وهي كما يلي:

أ - إنتاج بيانات إحصائية عامة عن المدونة، كحجمها، وعدد كلماتها دون تكرار، وعدد نصوصها.

ب - إنتاج قوائم التكرار والتكرار النسبي للمدونة للغوية كاملة، والتوزيع الإحصائي لتكرار الكلمات على أقسام المدونة، مثل: الأوعية والفترات الزمنية والمناطق إن كان مثل هذا التقسيم موجوداً. ويقصد بالتكرار عدد مرات ظهور الكلمة في المدونة، ويقصد بالتكرار النسبي نسبة ظهور الكلمة في المدونة مقارنة ببقية كلماتها. ويتم احتساب التكرار النسبي بقسمة تكرار الكلمة في المدونة على كامل حجم المدونة. ويستفاد من التكرار النسبي عند مقارنة التوزيع الإحصائي للكلمة بين مدونات مختلفة الأحجام.

ج - استخراج الكلمات الدليلية أو المميزة للمدونة اللغوية وذلك بمقارنة قوائم التكرار لكلمات لمدونة اللغوية مع قوائم التكرار لمدونة لغوية أخرى تسمى اصطلاحا بالمدونة المرجعية. وبالإمكان إجراء هذه المقارنة أيضاً على مستوى الوسوم النحوية والدلالية والصرفية متى ما كانت متوفرة والوسوم المستخدمة في المدونتين متطابقة.

د - إنتاج قوائم الكشاف السياقي للكلمة مناط البحث، والقصد من الكشاف السياقي هو استعراض جميع السياقات التي وردت فيها الكلمة داخل المدونة للكشف عن معانيها المختلفة، والكلمات التي تظهر بصحبتها في السياق، واختلاف المعنى من سياق لآخر، باختلاف الأوعية والفترات. ومن المناسب توفر إمكانية تحديد عدد الكلمات التي يتضمنها السياق قبل الكلمة مدار البحث وبعدها.

هـ - حسابات التصاحب اللفظي لكلمة معينة من خلال عدة معاملات إحصائية مثل مربع كاي Chi-Squired ومعامل كسب المعلومات Information Gain ومعامل المعلومات المتبادلة Mutual Information ومعامل الاحتمالية اللوغاريثمي Log Likelihood على سبيل المثال لا الحصر. ولا يشترط أن يكون الباحث على معرفة بالخلفية الرياضية

لهذه المعاملات الإحصائية، ويكفيه معرفة متى يستخدم هذه المعاملات وحدود استخدامها. وتسعى هذه المعاملات الإحصائية إلى الكشف عن مدى ارتباط الكلمة مع الكلمات الأخرى التي ظهرت معها في السياق، ولا يشترط في هذا الحساب أن تتوالى الكلمتان، بل أن تظهرا في سياق واحد حسبما يحدد الباحث حدود هذا السياق. ويمكن استخدام نفس المعاملات المذكورة سابقاً لحساب التلازم اللفظي بين كلمتين (أن تظهر الكلمة الأولى قبل الثانية مباشرة) مع تغيير طفيف في طريقة الحساب. ومتى ما كانت المدونة موسومة نحوياً أو دلالياً أو صرفياً فإنه يمكن حساب التصاحب أو التلازم النحوي أو الدلالي أو الصرفي.

تتوفر عدة برامج مجانية وأخرى بمقابل مادي للقيام بهذه الوظائف وغيرها، ومن أشهر البرامج المجانية لمعالجة المدونات برنامج أنتّ كونكْ (١٣) AntConc ولكنه لا يراعي اتجاه الكتابة العربية (من اليمين لليسار) أثناء عرض الكشاف السياقي. ولم يصمم نظام لمعالجة المدونات العربية يراعي خواصها ويحتوي كل ما ذكر أعلاه من وظائف سوى أداة معالجة المدونات العربية «غواص»(١٤) (الثبيتي وآخرون، ٢٠١٣ Al-Thubaity et al.) الذي تم تطويره في مدينة الملك عبدالعزيز للعلوم والتقنية. وعلى الرغم من أهمية ما توفره هذه الأدوات والبرامج من معلومات وبيانات إلا أن هذه المعلومات والبيانات ليست هي نهاية البحث اللغوي بل بدايته، ومهمة الباحث بعد ذلك هي الكشف عما تعنيه هذه البيانات من خلال تفحصه ودراسته لها.

مثال تطبيقي

يعرض هذا القسم مثالا تطبيقياً بسيطاً للخطوات التي سبق شرحها الخاصة بتصميم المدونات وبنائها. وقد تختلف معي عزيزي القارئ في التفاصيل التي سوف أضعها لمعايير التصميم أو لبناء المدونة، وهذا أمر طبيعي. ومرد هذا

الاختلاف – إن وجد – لسببين: الأول اختلاف الفهم للغرض من المدونة، والآخر هو اختلاف الخبرات والتجارب في هذا المجال، وعلى كل حال فإننا سوف نتفق على الكثير.

بداية سوف أحدد سؤال البحث أو الغرض الذي لأجله سوف أستخدم المدونة ثم أبحث عن مدى توفر مدونة لغوية بالإمكان استخدامها للإجابة عن سؤال البحث.

إن الغرض الذي حددته لاستخدام المدونة هو الكشف عن الفروقات في استخدام اللغة بين الكتّاب السعوديين والكاتبات السعوديات من خلال ما ينشر في الصحافة السعودية. وللأسف لم أجد أي مدونة لغوية يمكن أن تساعد في تحقيق هذا الغرض، ولاهتمامي بهذا الموضوع ومعرفتي باهتمام الكثيرين به سوف أضطر إلى بناء مدونة لغوية يمكنها الإجابة عن استفساراتي بهذا الخصوص. فيما يلي سوف أعرض للمعايير التي اخترتها لتصميم هذه المدونة:

معايير التصميم

أ –لغة المدونة: العربية الفصحى.

ب – طبيعة النصوص: النصوص المكتوبة.

ج – تاريخ النصوص: ٢٠١٤م.

د – المنطقة الجغرافية: المملكة العربية السعودية.

هـ – الوعاء: الصحف.

و – المجال: المقالات.

ز – الموضوعات: المقالات الاجتماعية، والثقافية، والدينية. وقد اقتصرت هنا على هذه المجالات الثلاثة لأنه من الممكن وجود كتابات لكلا الجنسين

فيها بينما يصعب هذا في مجالات أخرى مثل السياسية، والرياضية، والعلوم والتقنية.

ح – حجم العينة: النص كاملاً.

ط – جنس الكاتب: الذكور والإناث.

ي – حجم المدونة: ثلاثة ملايين كلمة، توزع بالتساوي بين المجالات وبين الجنسين مثلما هو موضح في الجدول (١)

المجال	الذكور	الإناث	المجموع
المقالات الاجتماعية	٥٠٠،٠٠٠	٥٠٠،٠٠٠	١،٠٠٠،٠٠٠
المقالات الثقافية	٥٠٠،٠٠٠	٥٠٠،٠٠٠	١،٠٠٠،٠٠٠
المقالات الدينية	٥٠٠،٠٠٠	٥٠٠،٠٠٠	١،٠٠٠،٠٠٠
المجموع	١،٥٠٠،٠٠٠	١،٥٠٠،٠٠٠	٣،٠٠٠،٠٠٠

جدول (١) توزيع كلمات المدونة على المجالات وجنس الكاتب

ك – معايير أخرى: قد يكون من المفيد تحديد معايير أخرى مثل المستوى التعليمي والفئة العمرية؛ لكن الوصول لهذه المعلومات وتحديدها بدقة صعب جداً.

ل – التمثيل والتوازن: كما تم توضيحه سابقاً فإن التمثيل والتوازن مفهومان مهمان ومترابطان، وقد حققت المدونةُ التوازنَ من خلال التوزيع المتساوي لعدد الكلمات بين المجالات وبين الجنسين. وعلى الرغم من غلبة عدد الكتّاب على الكاتبات في الصحافة السعودية وفي الصحف أيضاً إلا أن التساوي في توزيع عدد الكلمات والمجالات مهم جداً لأن القصد هو المقارنة بين كتابات الجنسين وليس أي نوع آخر من الكتابات، كما تحقق

التوازن من خلال حصر المجالات واقتصارها على المجالات التي من الممكن وجود كتابات صحفية للجنسين فيها بسهولة، وتحقق التمثيل من خلال تنوع المجالات ومن خلال اعتبار الصحف السعودية جميعها.

م - البيانات الأساسية: سوف أذكر هنا جميع البيانات الأساسية حتى ما هو بدهي منها، والغرض من هذا هو الاستفادة الكاملة من المدونة لاحقا فيما لو تم إضافتها لتكوين مدونات أخرى لنفس الغرض لكن تغطي فترات زمنية أخرى أو لتغطي بلداناً عربية أخرى. والمعلومات الأساسية المقترحة لنصوص لمدونتنا هذه هي: عنوان النص، اسم المؤلف، جنسية المؤلف، جنس المؤلف، وعاء النص، مجال النص، موضوع النص، تاريخ صدور النص باليوم والشهر والسنة، اسم الصحيفة، الرابط الإلكتروني للنص، تاريخ إضافة النص للمدونة اللغوية.

بناء المدونة

أ - حقوق الملكية الفكرية: كما أشرت سابقاً إلى أن المحافظة على حقوق الملكية الفكرية من أهم ما يجب التفكير فيه عند جمع نصوص المدونة، كما أن الرغبة في تعميم الاستفادة من المدونة وإتاحتها للباحثين الآخرين كمصدر لدراساتهم أو لضمها لمدونات أخرى يزيد من أهمية استئذان الصحف قبل البدء في جمع نصوص المونة اللغوية. بالطبع قد تأخذ هذه المرحلة بعض الوقت ولكن الإجابة بالقبول هي الغالبة، وعلى الباحث أن يدرك أهمية متابعة الموضوع والاتصال الدائم بالصحف وتوضيح فكرته ليحصل على الموافقة، كما يجب عليه ألا ينسى شكر الصحف وذكر موافقتها على ضم نصوصها لمدونته مع أهمية احتفاظه بنسخة من هذه الموافقة وإرفاقها مع المدونة.

ب – تحديد المصادر: الجدول (٢) يوضح مثالا للمصادر التي سيتم جمع نصوص المدونة منها ويمكن السير على منواله لتحديد بقية الصحف، مع العلم أن المعلومات الواردة في خانة موافقة الصحيفة هي معلومات افتراضية وليست حقيقية.

ج – الجمع: سيتم جمع نصوص المدونة يدوياً من مواقع الصحف بسبب أن المدونة صغيرة، كما أن تحديد جنس الكاتب آليا وتجهيز أحد البرامج لجمع نصوص كاتب معين يستغرق وقتاً طويلاً. ويمكن أن يقوم بهذا العمل عدة أشخاص يتولى كل منهم جمع النصوص من صحيفة معينة، ويجب الحرص على تنوع الكتّاب والكاتبات وعدم الاقتصار على كتّاب وكاتبات بعينهم.

د – الترميز والتسمية والحفظ: يتم حفظ النصوص في ملفات نصية بصيغة TXT بترميز UTF8، ويتم حفظ نصوص الكتاب في مجلد باسم SNP_M(١٥)، والكاتبات في مجلد باسم SNP_F(١٦). بالنسبة لتسمية الملفات سوف نقسم الاسم إلى ٥ أجزاء كما هو موضح في الجدول (٣)

م	الصحيفة	العنوان	الموقع على الشابكة	موافقة الصحيفة
١	الرياض	المملكة العربية السعودية – الرياض، حي الصحافة أول طريق القصيم ص.ب ٨٥١ الرياض ١١٤٢١ سنترال: ٢٩٩٦٠٠٠، فاكس: ٤٨٧١٠٧٠	http://www.alriyadh.com/	إعداد الخطاب

٢	مكة	مكتب صحيفة مكة الرئيسي في مكة المكرمة: صندوق بريد: ٥٨٠٣، الرمز البريدي: ٢١٩٥٥ تليفون: ٩٦٦١٢٥٢٠٦٧٧٦، فاكس: ٩٦٦١٢٥٢٠٣٠٥٤	http://www.makkahnewspaper.com/	تم إرسال الخطاب
٣	اليوم	السنترال: ٩٦٦ ٣ ٨٥٨٠٨٠٠، الفاكس: ٩٦٦ ٣ ٨٥٨٨٧٧٧ الرقم المجاني: ٨٠٠٦١٢١٢١٢ ص.ب. ٥٦٥ الدمام ، ٣١٤٢١، المملكة العربية السعودية mail@alyaum.com	http://www.alyaum.com/	تم التواصل إلكترونياً، في انتظار الرد
٤	الوطن	أبها – مدينة سلطان ، طريق المطار التحرير– هاتف مجاني (٨٠٠٧٥٤٠٠٠٧)، سنترال ٢٢٧٣٣٣٣ فاكس التحرير ٢٢٧٣٧٥٦ ، ص.ب. ١٥١٥٥	http://www.alwatan./com.sa	تمت الموافقة

جدول (٢) مصادر المدونة

القسم	الدلالة	القيم الممكنة
١	السنة	٢٠١٤
٢	جنس الكاتب	١ للكتّاب، ٢ للكاتبات
٣	الموضوع	CUL الثقافية، REL الدينية، SOC الاجتماعية
٤	الصحيفة	يأخذ عدداً صحيحاً من خانتين كل عدد منها يدل على صحيفة معينة.
		الرياض:٠١، اليوم:٠٢، مكة:٠٣، الوطن:٠٤ وهكذا لبقية الصحف
٥	التسلسل في المدونة	قم يتكون من أربع خانات يدل على تسلسل النص في المدونة

جدول (٣) تسمية ملفات المدونة

لو نظرنا لاسم الملف التالي ٢٠١٤-٢-CUL-٠٤-٠٠٩١ لعرفنا أنه لكاتبة، وأن موضوعه ثقافي، وأنه من صحيفة الوطن السعودية، وأن تسلسله في المدونة هو ٩١. وسوف تحفظ جميع المعلومات الأساسية للمدونة اللغوية في قائمة كما هو موضح في الجدول (٤)

اسم الملف	العنوان	جنس الكاتب	الكاتب	الموضوع	الصحيفة	رابط النص
2014-2-CUL-04-0001	إلى الساخرين من التراث!!	أنثى	ملحة عبدالله	ثقافي	الوطن	http://www.alwatan.com.sa/Articles/Detail.aspx?ArticleId=22691
2014-1-SOC-02-0002	تحصين المجتمع الجامعي لمواجهة المتغيرات	ذكر	عادل رشاد غنيم	ديني	اليوم	http://www.alyaum.com/article/4013773

جدول (٤) المعلومات الأساسية للمدونة اللغوية

الخاتمة

المدونات ما هي إلا بيانات لا تختلف عن أي بيانات أخرى تجمع لإجراء الدراسات العلمية، وهي عينة من اللغة وليست اللغة كلها، وبحسب وضوح المنهج المتبع وانضباطه في جمع المدونات يمكن الاعتماد والوثوق بنتائج الدراسات القائمة عليها؛ فالهدف من استخدام المدونات في دراسة اللغة هو الكشف عن الأنماط الشائعة في اللغة المستخدمة الذي قد يتطابق أو يختلف عما نعرفه عنها معيارياً.

وقد شرح هذا المبحث بشكل مختصر ومبسط خطوات إجرائية، تصلح إطاراً عاماً يمكن اتباعه عند تصميم المدونات وبنائها، والأدوات المهمة الرئيسية لمعالجة هذه المدونات. حيث وضح القسم الثاني من هذا الفصل أهم ما يجب النظر إليه عند تصميم المدونات مثل لغة المدونة، وطبيعة نصوصها، وتاريخ هذه النصوص، ومكان صدورها، والأوعية التي ظهرت فيها، والمجالات المختلفة التي تغطيها، كما ذكرنا في هذا القسم النقاط التي تراعى في تحديد حجم المدونة وحجم العينات التي تؤخذ للوصول للحجم المطلوب مع الإشارة إلى المعايير الأخرى التي قد يكون الباحث بحاجة لأخذها بالحسبان، كما تطرق القسم الثاني إلى مفهومين مهمين في المدونات يحددان أهميتها، هما التمثيل والتوازن.

وتطرق القسم الثالث من هذا المبحث إلى النقاط التي ينبغي الاهتمام بها عند جمع نصوص المدونة استرشاداً بمعايير التصميم المشروحة في القسم الثاني، فتناول حقوق الملكية الفكرية، وتحديد مصادر جمع النصوص، وكيفية جمع النصوص مع التطرق إلى بعض المسائل التقنية عند الانتهاء من الجمع وهي ترميز ملفات نصوص المدونة وتسميتها وحفظها وإضافة معلومات مساعدة في تعزيز الفائدة من المدونة من خلال تحشيتها بمعلومات إضافية مختلفة.

وتطرق القسم الرابع باختصار إلى أهم الأدوات التي تساعد في تحليل المدونات ودراستها، فتطرق إلى خمس وظائف يمكن أن تساعد بشكل أساسي في هذا الشأن هي الإحصاءات العامة عن المدونة بمجملها، وقوائم التكرار، والكلمات المميزة للمدونات، والكشاف السياقي، والتصاحب اللفظي، ثم تطرق القسم السادس إلى مثال تطبيقي لما سبق شرحه.

ومع أهمية اتباع هذه الخطوات والإجراءات لجمع عينة من اللغة تكون ممثلة للغة أو إحدى صورها مجال الدراسة فإن الجهد يجب ألا يتوقف عند هذا الحد؛ فتصميم المدونة ونصوصها يجب أن يخضعا للمراجعة والتقويم المستمر متى ما استدعى الأمر ذلك، مع إتاحتها للباحثين الآخرين متى ما كان ذلك ممكناً.

الحواشي

(٧) http://www.sketchengine.co.uk/

(٨) المد الذي يضاف في وسط الكلمة بغرض تصفيف الكلمة ومحاذاتها

(٩) http://www.info.gov.sa/copyrights/SectionDetails.aspx?id=6

(١٠) http://bootcat.sslmit.unibo.it/

(١١) http://www.tei-c.org/index.xml

(١٢) http://www.xces.org/

(١٣) http://www.antlab.sci.waseda.ac.jp/software.html

(١٤) http://sourceforge.net/projects/ghawwasv4

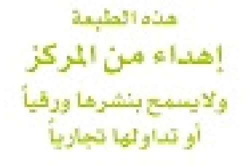
(١٥) SNP_M:Saudi NewsPapers_Males الصحف السعودية___ الذكور

(١٦) SNP_F: Saudi NewsPapers_Females الصحف السعودية___ الإناث

المراجع

الثبيتي، عبدالمحسن. المدونات العربية، التحليل الصرفي والتوسيم النحوي. محاضرة بمركز اللغويات التطبيقية. جامعة الإمام محمد بن سعود الإسلامية. الرياض. ٧ مايو ٢٠١٤م. (٢٠١٤. http://www.slideshare.net/ althubaity/ss-34381929)

Ahmad, Khurshid. «Being in Text and Text in Being: Notes on representative texts.» Incorporating Corpora: The linguist and the translator (2008): 60-94.



Alansary, Sameh, Magdy Nagi, and Noha Adly. «Building an International Corpus of Arabic (ICA): progress of compilation stage.» 7th International Conference on Language Engineering, Cairo, Egypt, 5–6 December 2007. 2007.

Alfaifi, A. Y. G., Eric Atwell, and I. Hedaya. «Arabic learner corpus (ALC) v2: a new written and spoken corpus of Arabic learners.» (2014).

Alkanhal, Mohamed I., et al. «Automatic Stochastic Arabic Spelling Correction With Emphasis on Space Insertions and Deletions.» Audio, Speech, and Language Processing, IEEE Transactions on 20.7 (2012): 2111-2122.

Al-Sulaiti, Latifa, and Eric Steven Atwell. «The design of a corpus of contemporary Arabic.» International Journal of Corpus Linguistics 11.2 (2006): 135-171.

Al-Thubaity, A. O. (2014).A 700M+ Arabic corpus: KACST Arabic corpus design and construction. Language resources and evaluation. DOI 10.1007/s10579-014-9284-1

Al-Thubaity, Abdulmohsen, et al. «New Language Resources for Arabic: Corpus Containing More Than Two Million Words and a Corpus Processing Tool.»Asian Language Processing (IALP), 2013 International Conference on. IEEE, 2013.

Alzahrani, Salha M. «Building, Profiling, Analysing and Publishing an Arabic News Corpus Based on Google News RSS Feeds.» Information Retrieval Technology. Springer Berlin Heidelberg, 2013. 488-499.

Atkins, Sue, Jeremy Clear, and Nicholas Ostler. «Corpus design criteria.»Literary and linguistic computing 7.1 (1992): 1-16.

Biber, Douglas, Susan Conrad, and Randi Reppen. Corpus linguistics: Investigating language structure and use. Cambridge University Press, 1998.

Biber, Douglas. «Representativeness in corpus design.» Literary and linguistic computing 8.4 (1993): 243-257.

Crystal, David. An encyclopedic dictionary of language and languages. Middlesex, UK: Blackwell, 1992.

Jakubíček, Miloš, et al. «The TenTen Corpus Family.» Proc. Int. Conf. on Corpus Linguistics. 2013.

Khoja, Shereen. «An RSS Feed Analysis Application and Corpus Builder.»Proceedings of the Second International Conference on Arabic Language Resources and Tools. 2009.

Khorsheed, Mohammad S., and Abdulmohsen O. Al-Thubaity. «Comparative evaluation of text classification techniques using a large diverse Arabic dataset.» Language resources and evaluation 47.2 (2013): 513-538.

Leech, Geoffrey. «New resources, or just better old ones? The Holy Grail of representativeness.» Language and Computers 59.1 (2006): 133-149.

McEnery, Tony, Richard Xiao, and Yukio Tono. Corpus-based language studies. London: Routledge, 2006.

Meyer, Charles F., ed. English corpus linguistics: An introduction. Cambridge University Press, 2002.

Sinclair, John. «Corpus and text-basic principles.» Developing linguistic corpora: A guide to good practice (2005): 1-16.

Sinclair, John. Corpus, concordance, collocation. Oxford University Press, 1991.

Teubert, Wolfgang, and Anna Cermáková. Corpus linguistics: A short introduction. Continuum, 2007.

المبحث الرابع

لسانيات المدونات:
نماذج وتطبيقات في لغة الصحافة العربية

عقيل بن حامد الشمري و عبدالمحسن بن عبيد الثبيتي
جامعة الملك سعود
مدينة الملك عبدالعزيز للعلوم والتقنية
alaqeeel@ksu.edu.sa
aalthubaity@kacst.edu.sa

مقدمة

الهدف الأساسي لهذا المبحث هو تقديم نماذج تطبيقية، وعينات توضح طرق استخدام «لسانيات المدونات»، وسبل استثمار إمكاناتها الحاسوبية في رصد وتحليل الظواهر اللغوية. وأما المادة اللغوية المستخدمة في التحليل فهي «العربية الفصحى المعاصرة» المستخدمة في لغة الصحافة المكتوبة في كافة الأقطار العربية. ولقد اخترنا هذا الموضوع لربط لسانيات المدونات كمنهجية في الدراسة والتحليل بالإشكال النظري المتعلق ببعض ظواهر العربية المعاصرة لكي يتسنى لنا مناقشة ما يتعلق بذلك من قضايا ومسائل مع التمثيل المناسب المعتمد على «لسانيات المدونات». وينقسم المبحث إلى أربعة أقسام رئيسية. فبعد أن نتناول مفهوم «الفصحى المعاصرة» وما يتعلق به من إشكالات نظرية ومنهجية في القسم الأول، سننتقل في القسم الثاني إلى وصف المدونة وأداة التحليل المستخدمة في الدراسة الحالية. أما القسم الثالث فيستعرض أهم النتائج المستخلصة التي توضح طرق استخدام المدونات في التحليل اللغوي. والقسم الرابع يتضمن مناقشة عامة وموجزة تربط النتائج المستخلصة بالإطار النظري، وتوضح الإضافة المنهجية التي تقدمها لسانيات المدونات، وما تفتحه من آفاق جديدة في مجال البحث اللغوي.

الإطار النظري

٤. ١ الفصحى المعاصرة بين الرفض والقبول وضعف البحث العلمي

اللغة العربية ذات امتداد تاريخي طويل، وانتشار جغرافي واسع. وقد مرت بمراحل متعددة ومختلفة (انظر مثلاً: فريمان، ٢٠١٣). ومن ضمنها مرحلة العصر الحديث، حيث شهدت اللغة العربية منذ مطلع القرن التاسع عشر الميلادي تغيرات واسعة، بسبب تأثيرات حركة «النهضة العربية الحديثة» (١٨)،

بعد فترة طويلة من الضعف والركود عرفت في تاريخ آداب العربية بـ«عصور الانحطاط والتدهور» (للمزيد حول هذا الموضوع، انظر: بيلكين، ١٩٧٣؛ فرستيغ، ٢٠٠٣). وقد درج كثير من الباحثين والدارسين، سواء من العرب (مثل: بدوي، ٢٠١٢؛ عبدالعزيز، ١٩٩٨؛ الحمزاوي، ١٩٨٦) أو من غيرهم (مثل: فرستيغ، ٢٠٠٣؛ Bateson، ٢٠٠٣؛ Ryding، ٢٠٠٥)، على اعتبار العربية الفصحى المستخدمة في العصر الحديث مستوى لغوياً متميزاً، ينعتونه بـ«الفصحى المعاصرة» في مقابل «فصحى التراث» من ناحية، والعاميات أو اللهجات المنطوقة والدارجة في المجتمعات العربية من ناحية أخرى.

ومع أن «الفصحى المعاصرة» عند أصحاب هذا الاتجاه امتداد للعربية في كل عصورها وأزمانها، إلا أن لها طابعها الخاص الذي يشتمل على بعض الظواهر والسمات المميزة(١٩). وهذه الظواهر متنوعة ومتداخلة، ولكنها تعود في معظمها إلى ثلاثة جوانب أساسية:

أ- جوانب تركيبية: تتعلق بنوعية الجمل المستخدمة (خلف، ٢٠١١)، واستحداث تراكيب جديدة غير معهودة في العربية (عبدالعزيز، ١٩٩٨؛ الزعبي، ٢٠٠٦)، بالإضافة إلى بعض الاختلافات في تعدية الأفعال واستخدام حروف الجر (السامرائي، ١٩٩٥؛ حمادي، ١٩٩٩؛ البلداوي، ٢٠١٢؛ نصار وحماد، ٢٠١٤)، والترخص أو المرونة في بعض القضايا المتعلقة بالمطابقة والرتبة (عبدالعزيز، ١٩٩٩).

ب- جوانب معجمية: تتعلق بتوليد واشتقاق أبنية وصيغ جديدة (ستتكفيتش، ١٩٨٥؛ القضماني وعبدالقادر، ٢٠٠٤)، وظهور تعابير مستحدثة (فايد، ٢٠٠٣)، بالإضافة إلى غلبة المفردات الحديثة، ذات الارتباط بالحياة الاجتماعية المعاصرة، والتي استحدثت إما عن طريق الاقتراض، أو التوليد، أو العدول عن الدلالات القديمة إلى دلالات أخرى جديدة (ستتكفيتش، ١٩٨٥).

ج – جوانب أسلوبية: تتعلق بالترسل، والميل نحو الوضوح والسهولة، والتخلص من المحسنات البديعية (البعلبكي، ١٩٨٨)، واستحداث أساليب جديدة متأثرة بالترجمة (ستتكفيتش، ١٩٨٥؛ عصفور، ٢٠٠٧؛ أبو الهيجاء، ٢٠١٠)، بالإضافة إلى تطور أجناس الكتابة النثرية وموضوعاتها (القاعود، ٢٠٠٨).

وأغلب الدراسات التي تطرقت إلى الظواهر المتعلقة بهذا الموضوع تعتمد أصلاً على جهود مجمع اللغة العربية في القاهرة وقراراته في الألفاظ والأساليب المستجدة (انظر مثلاً: مجمع اللغة العربية، ١٩٨٣؛ ١٩٨٩).

ورغم كثرة ما كتب في هذا الموضوع، إلا أن مفهوم «الفصحى المعاصرة» لا يزال مفهوماً ملتبساً، يعاني الكثير من الاضطراب المنهجي والاصطلاحي (للوقوف على بعض الإشكالات الاصطلاحية حول مفهوم «الفصحى المعاصرة»، انظر: عبدالكريم، ٢٠٠٨). فرغم أهمية الإشارات والملاحظات المتعلقة بالجوانب التي طالها شيء من التغير والاختلاف في الفصحى في العصر الحديث، إلا أن تلك الملاحظات لا تزال ملاحظات جزئية، ذات طبيعة مجملة. فنوعية التغيرات فيما يسمى بـ«الفصحى المعاصرة»، وحجمها، ومدى أثرها، وطبيعة مخالفتها لأساليب العربية ومستوياتها السابقة كل ذلك لم يدرس دراسة وافية، ولم يفهم الفهم الكافي بعد (بشر، ١٩٩٩، خلف، ٢٠١١).

وفي رأينا أن أسباب القصور في تناول هذا الموضوع تعود إلى جانبين، أحدهما نظري، والآخر منهجي. وبالنسبة لقصور الجانب النظري، فإن كثيرا من الأحكام في هذا الصدد لا تستند إلى نظرية واضحة، وشاملة، وموحدة في تحديد المستويات اللغوية، وما يميز كل مستوى عن الآخر. وما كتب في هذا الموضوع (انظر مثلا: بدوي، ٢٠١٢، السوسوة، ٢٠٠٢) لا يعدو أن يكون محاولات أولية غير مكتملة، خاصة في ظل الواقع المعقد للسانيات الاجتماعية العربية الذي يشتمل على العديد من الظواهر والمستويات المتداخلة (انظر مثلا: Bassiouney،

٢٠٠٩). ولذا فإن مفهوم االفصحى المعاصرة، رغم أنه قد يكون له مدلولات صحيحة، إلا أنه لا يزال من الناحية النظرية مفهوما غير مكتمل، خاصة أن مجمل مراحل التطور التاريخي للعربية لا تزال غير مفهومة بشكل كاف (انظر مثلا: فريمان، ٢٠١٣).

وأما بالنسبة لقصور الجانب المنهجي، فرغم وجود عدد من المحاولات الجادة والجيدة لتتبع الظواهر والخصائص المستجدة في العربية المعاصرة (مثل: ستتكيفتش، ١٩٨٥)، إلا أن كثيرا مما يكتب في هذا الموضوع لا يستند إلى منهج واضح(٢٠). فأغلب ما يكتب في هذا الموضوع عبارة عن ملاحظات مبنية على شواهد وأمثلة جزئية ومحدودة (انظر مثلا: السامرائي، ١٩٩٥، عبدالعزيز، ١٩٩٨)، ولا يقوم على منهج اختباري وإحصائي لظواهر الاستعمال الفعلي يقارن مستويات الاستخدام اللغوي، ويسمح بتراكم نتائج البحث ومقارنة الدراسات.

٤. ٢ الأنواع اللغوية في «الفصحى المعاصرة»: لغة الصحافة نموذجاً

إن جوانب القصور في مفهوم «الفصحى المعاصرة» لا تقتصر على ما ذُكِر أعلاه، بل تتعدى ذلك إلى إغفال التنوعات المتعلقة بمستويات ومجالات الاستخدام اللغوي المختلفة داخل العربية. فمفهوم «الفصحى المعاصرة» مفهوم واسع جدا، وبسبب القصور النظري والمنهجي الذي ذكرناه آنفا فإنه لا يراعي اختلاف السجلات وضروب الاستخدام اللغوي(٢١)، وما بينها من فروقات. فلِسَعَة مفهوم الفصحى المعاصرة فإنه يشمل لغة الصحافة والكتب والرسائل العلمية، والروايات والقصص الأدبية، والمكاتبات الإدارية والرسمية، وغيرها. ويشمل أيضا أنماطا مختلفة من التعبير كالنمط العلمي، والنمط الأدبي، والنمط الإعلامي، وغيرها. ولذا فإننا نعتقد أن البداية الصحيحة لدراسة «الفصحى المعاصرة» لا تكون بدراسة الأمثلة والشواهد الجزئية، بل بالانطلاق من الأنواع

اللغوية المختلفة لتوضيح ما بينها من مشتركات، وما يختص به كل نوع من سمات وخصائص مميزة.

ورغم كثرة الأنواع اللغوية وتعددها، فإن الصحافة المكتوبة تعتبر أحد أبرز مجالات الاستخدام اللغوي في المجتمعات المعاصرة. وبالنسبة للعربية، فإن صعودها وانبعاثها في العصر الحديث ابتدأ أصلا بنشوء الصحافة الحديثة (بيلكين، ١٩٧٣)، التي ظلت الوعاء الأساسي الحامل للفصحى المعاصرة في المجتمع، وأبرز عوامل توسيعها وانتشارها (انظر مثلا: كنون، ١٩٨٣؛ عبدالعزيز، ١٩٨٧؛ البعلبكي، ١٩٨٨؛ الحمزاوي، ٢٠٠٣). ومن ميزات الصحافة أنها تحتوي على عدد متنوع من الموضوعات والمجالات من ثقافية وعلمية ودينية واجتماعية وغيرها، مما يجعلها تحتوي على أنماط متنوعة من التعبير والاستخدام اللغوي.

ومع ذلك لم تحظ الصحافة العربية بالدراسة اللغوية التي تستحقها. فأغلب ما كتب ويكتب في هذا الموضوع يندرج ضمن ما يسمى بـ«حركة التصحيح اللغوي» (حمادي، ١٩٨١) الذي يقوم على تتبع ما يرد في وسائل الإعلام من «أخطاء» وتصويبها (انظر مثلا: عمر، ١٩٩٣)، على ما يعتري ذلك من اضطراب في معايير التصحيح، واختلاف في آراء المصححين (الحمزاوي، ٢٠٠٣؛ السامرائي، ٢٠٠٠). ولا يوجد إلا عدد محدود جداً من الدراسات التي بدأت مؤخرا تستخدم الوصف والتحليل في دراسة أنماط ومكونات وأساليب التعبير اللغوي في الصحافة العربية (للاطلاع على بعض الأمثلة، انظر: عبدالعزيز، ١٩٩٨؛ سمبس، ٢٠٠٦). ورغم أهمية هذه الدراسات الوصفية والتحليلية إلا أنها لا تزال في معظمها تستخدم الأسلوب التقليدي الذي يعتمد على الفرز اليدوي في استخراج البيانات، واستعراضها، وتصنيفها (للاطلاع على بعض النماذج، انظر: فضل، ٢٠١٠) بعيداً عما أنجزته لسانيات المدونات من تقدم كبير في مناهج الإحصاء والدراسة والتحليل. وهذه النقطة بالذات هي ما ستحاول الدراسة الحالية التعامل معه كما سنشرح في الجزء التالي.

الدراسة الحالية: لسانيات المدونات وإمكانات البحث المتاحة

الهدف الأساسي لهذا الفصل هو استخدام لسانيات المدونات ومناهجها في الدراسة والبحث اللغوي لحل بعض الإشكالات النظرية والمنهجية التي ذكرت في القسم السابق. فلسانيات المدونات تمثل منهجية اختبارية/empirical بديلة في البحث اللغوي تتجاوز التتبع العشوائي لظواهر جزئية ومتفرقة إلى التحليل المنظم لأنماط وأشكال الاستخدام الفعلي للغة بالاعتماد على الإمكانات والأدوات الحاسوبية في الجمع والإحصاء والتصنيف واسترجاع البيانات (العصيمي، ٢٠١٣). ولذا فإنها توفر آلية مناسبة ومفيدة لدراسة العربية في العصر الحديث وما تحتوي عليه من أنواع ونماذج مختلفة ومتنوعة. ورغم أن لسانيات المدونات بذاتها لا تمثل نظرية في اللغة(٢٢)، ولذا فإنها لا تتعامل مع الإشكال النظري بشكل مباشر، إلا أنها تؤدي إلى استخلاص نتائج اختبارية موثوقة ومنظمة قابلة للتكرار والتراكم والمقارنة والتعميم، وهو ما يساعد على بناء النظريات واختبارها وتهذيبها على المدى الطويل.

ورغم أن مفهوم «الفصحى المعاصرة» وما يتعلق به من إشكالات وقضايا يتجاوز حدود الدراسة الحالية والمساحة المتاحة، إلا أن المقصود هو تقديم نماذج وأمثلة لاستخدام المدونات في البحث اللغوي المرتبط بهذا الموضوع. ولذا فإننا سنكتفي في الدراسة الحالية باستعراض بعض العينات الممثلة دون الدخول في تفاصيل كثيرة ومتشعبة، مع توضيح ارتباطات النتائج المستخلصة بما أثاره الإطار النظري من قضايا حول «الفصحى المعاصرة». فاستخدام المدونات يتيح أنواعا متعددة ومختلفة من التحليل والدراسة(٢٣)، إما بالانطلاق من فرضيات وتصورات نظرية مسبقة لفحصها والتحقق منها، أو بالاعتماد على رصد واستكشاف الظواهر البارزة في البيانات بدون افتراضات مسبقة.

وللتمثيل على ما نريد، فإننا سنستخدم في الدراسة الحالية كلتا الطريقتين، لا سيما أن الدراسات السابقة التي استعرضنا بعضها في الجزء السابق توفر بعض المعطيات التي يمكن اعتبارها افتراضات أولية حول ظواهر الاستخدام في «الفصحى المعاصرة»، خاصة قرارات المجمع وجهوده في دراسة الألفاظ والأساليب المستجدة في العربية (مجمع اللغة العربية، ١٩٨٩، وانظر أيضاً: ستتكفيتش، ١٩٨٥).

بيانات الدراسة وأدوات التحليل

٤,١. المدونة

سوف نستخدم في دراستنا هذه المدونة اللغوية للصحف العربية لعام ٢٠١٢ (الثبيتي وآخرون Al-Thubaity et al.، ٢٠١٣). وتغطي هذه المدونة اللغوية عينة من الكتابات الصحفية من تسعة عشر بلداً عربياً موزعة على ستة مجالات. ويزيد الحجم الإجمالي للمدونة اللغوية للصحف العربية لعام ٢٠١٢ عن ٢,١ مليون كلمة(٢٤). ورغم صغر حجمها نسبياً (انظر المبحث الثالث الجزء المتعلق بحجم المدونات)، إلا أن المدونة اللغوية للصحف العربية لعام ٢٠١٢ كافية من وجهة نظرنا للوفاء بغرض دراستنا الذي أشرنا إليه آنفاً حيث إنها تعتبر مدونة لغوية متوازنة وممثلة للغة الصحافة العربية لسببين رئيسيين. أولهما أنها ضمت نصوصاً من صحافة جميع البلدان العربية بشكل متساوٍ تقريباً(٢٥). وثانيهما أنها ضمت نصوصا من أغلب الموضوعات المشتركة في الصحف العربية بشكل عام. الجدول (٥) يوضح توزيع الكلمات على البلاد العربية التي شملتها المدونة، كما يوضح الجدول (٦) توزيع الكلمات على مجالات المدونة.

البلد	عدد النصوص	عدد الكلمات
المغرب	٢١٧	١٤٤،٩٧٨
البحرين	٢٠٦	١٣٩،٨٥٥
الإمارات	١٧٧	١٣٨،٠٣١
الأردن	٢٠٨	١٣٣،٢٩٤
السودان	١٤٦	١٣٢،٢٣٨
السعودية	١٤٩	١٢٧،٧٢٨
العراق	١٥٨	١٢٦،٩٧٦
قطر	١٧٤	١٢٦،٤٢٤
عمان	١٣٩	١٢٦،٣٧١
فلسطين	١٨٧	١٢٥،١٩٨
الكويت	١٦١	١٢٤،٩٣٥
مصر	١٩٢	١٢٣،٨٦٤
اليمن	١٥٣	١٢٢،٦٧١
الجزائر	٢١٥	١٢٠،١١٢
تونس	١٣٨	١٠٣،٩٠٦
لبنان	١٢٥	١٠٣،٤٦٧
سوريا	١٠١	٨٣،٧٢٩
موريتانيا	٣٣	٥٠،٧٥٩
ليبيا	٣١	٢٩،٨٣٩
المجموع	٢،٩١٠	٢،١٨٤،٣٧٥

جدول (٥) توزيع كلمات المدونة على البلدان العربية

المجال	عدد النصوص	عدد الكلمات
الثقافي	٤٦٤	٤٠٠،٨٤٩
السياسي	٤٧٤	٣٨٠،٤٦٧
الرياضي	٥٣٤	٣٧١،٩٣٥
الديني	٤٣٠	٣٦٠،٦٧٥
الاقتصادي	٥٣٤	٣٦٠،١٩٠
العلوم والتقنية	٤٧٤	٣١٠،٢٥٩
المجموع	٢،٩١٠	٢،١٨٤،٣٧٥

جدول (٦) توزيع كلمات المدونة على المجالات

٢,٤. أداة التحليل

اخترنا أداة معالجة المدونات العربية «غواص» الذي تم تطويره في مدينة الملك عبدالعزيز للعلوم والتقنية من أجل إجراء عمليات البحث والمقارنة بين أقسام المدونة اللغوية المستخدمة في الدراسة الحالية. ويتوفر في غواص العديد من الإمكانات المساعدة في دراسة المدونات وبالأخص العربية منها. وسوف نتطرق لأهم الوظائف التي سوف تساعدنا في هذه الدراسة. وهذه الوظائف كما يلي:

– معلومات عامة عن المدونة اللغوية تشمل عدد النصوص، وعدد الكلمات الكلي (حجم المدونة) وعدد الكلمات بدون تكرار.

– قوائم التكرار والتكرار النسبي وتكرار النصوص والتكرار النسبي للنصوص التي وردت فيها الكلمات والمتواليات اللفظية.

– الكشاف السياقي للكلمة أو المتواليات اللفظية مع إمكانية تغيير عدد الكلمات السابقة واللاحقة ما بين كلمة إلى خمس عشرة كلمة.

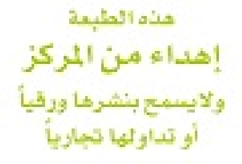
– تحديد الكلمات المميزة للمدونة عند مقارنتها بمدونة أخرى باستخدام طرق إحصائية مختلفة.

– البحث عن كلمة حسب رسمها الإملائي أو بحسب جذعها.

– البحث عن كلمة أو متوالية لفظية حسب موضع حرف أو حروف منها.

– إمكانية إهمال قائمة معدة مسبقاً من الكلمات أو المتواليات فلا تظهر في نتائج قوائم التكرار.

– إمكانية حصر البحث على قائمة معدة سلفاً من الكلمات أو المتواليات اللفظية.



- إمكانية حفظ نتائج جميع الوظائف السابقة للمدونة كاملة أو حسب توزيعها على الأقسام أو حسب الملفات كلاً على حدة.

نتائج الدراسة

٥.أ. الكلمات الشائعة والأكثر تكراراً

التوزيع الإحصائي لتكرار الكلمات في المدونة الذي يتضمن استخراج الكلمات الشائعة، وعدد مرات ورودها، ونسبة تكرارها من أبرز وأهم العمليات المستخدمة في تحليل المدونات. ويعرض الجدول رقم (٧) قائمة بالكلمات المئة الأكثر تكراراً في المدونة كلها (٢٦). وكما يظهر فإن مجموع التكرار النسبي لهذه الكلمات المئة يعادل تقريبا ربع كلمات المدونة، وهي نسبة كبيرة جداً، تدل على أهمية هذه الكلمات، ومدى شيوعها وتكررها. وبالنظر إلى هذه الكلمات نجد أن الغالبية العظمى منها تنتمي إلى فئة «الكلمات الوظيفية»، وهي فئة مغلقة تشتمل على «الأدوات» أو «حروف المعاني»(٢٧)، وما يشابهها مثل الضمائر، وأسماء الإشارة، والأسماء الموصولة.

#	الكلمة										٪
١٠	في	من	على	أن	إلى	التي	عن	ما	الذي	لا	٪١١,٨٥
٢٠	مع	هذا	الله	أو	هذه	ان	و	بين	كان	الى	٪٣,١١
٣٠	كل	ذلك	بعد	كما	خلال	لم	هو	إن	حيث	عليه	٪٢,١١
٤٠	وفي	ولا	وهو	قد	حتى	قبل	ومن	له	العام	كانت	٪١,٤٢
٥٠	هي	قال	بن	أنه	غير	أي	إلا	بعض	علي	فيها	٪١,٢٢
٦٠	وقد	محمد	وقال	بها	به	ثم	عام	العالم	فيه	الذين	٪١,٠٣
٧٠	أكثر	يكون	العربية	يمكن	اليوم	وهي	هناك	لها	لكن	رئيس	٪٠,٩٢
٨٠	تلك	مثل	العربي	إذا	منها	فإن	منذ	بشكل	بل	الأول	٪٠,٨٠
٩٠	دون	عند	ولكن	عبد	الناس	العمل	فقد	تم	أيضا	بأن	٪٠,٧٤
١٠٠	عليها	الرئيس	تكون	خاصة	حول	أخرى	مليون	وما	عدد	أما	٪٠,٦٨
									مجموع التكرار النسبي		٪٢٣,٨٨

الجدول (٧) الكلمات المئة الأكثر تكراراً في المدونة

ورغم أن القائمة تضم عددا من الأسماء، إلا أن أغلبها ينتمي إلى فئة «الأسماء المبهمة»، مثل «بين، كل، بعد، خلال، حيث، قبل، غير، بعض، مثل، دون، عند، حول». والأسماء المبهمة، وإن كانت من قبيل كلمات المحتوى، إلا أنها قريبة من الكلمات الوظيفية وتشابهها في كثير من سماتها وخصائصها(٢٨)، مثل الدلالة على معنى وظيفي كالظرفية، وعدم التصرف، والاحتياج إلى الإضافة أو ما يشبهها في تتميم معناها وتعيين مقصودها. وأما الأسماء المعينة فقليلة، وكذلك الأفعال، وهي أقل أنواع الكلمات في القائمة، فلم يرد فيها إلا ثلاثة أفعال فقط: «قال، يمكن، تم».

ولا يقتصر الفرق بين الكلمات الوظيفية وكلمات المحتوى على كثرة التكرار في الأولى وقلته في الثانية، بل يوجد بينهما فرق آخر من ناحية نسبة عدد النصوص التي يتكرران فيها. فالتكرار النسبي للنصوص التي وردت فيها كلمات المحتوى أقل بكثير من تكرار النصوص التي وردت فيها الأدوات والكلمات الوظيفية، كما توضح الأمثلة في الجدول (٨).

الكلمة	تكرار الكلمة		تكرار النصوص	
	التكرار	نسبة التكرار	التكرار	نسبة التكرار
الله	٨٠٨٣	٠,٣٧	٩٢٧	٣١,٨٥
أو	٧٨٧٢	٠,٣٦	١٧٦٩	٦٠,٧٩
هي	٢٨٢٨	٠,١٣	١٣٣١	٤٥,٧٣
قال	٢٨٠٥	٠,١٣	٩٦٥	٣٣,١٦
بن	٢٧٩٦	٠,١٣	٦٠٩	٢٠,٩٢
أنه	٢٧٧٣	٠,١٣	١٣٣٨	٤٥,٩٧

الجدول (٨) مقارنة التكرار النسبي للنصوص لكلمات المحتوى والكلمات الوظيفية المقاربة لها في التكرار

وكما يظهر في الجدول (٨)، فإن لفظ الجلالة (الله)، على سبيل المثال، هو أكثر الأسماء تكراراً في المدونة، ولكن مقارنته بأداة العطف (أو) المقاربة له في عدد مرات التكرار تبين أن نسبة تكرار النصوص التي وردت فيها (أو) أكثر بكثير من نسبة النصوص التي ورد فيها لفظ الجلالة. وينطبق الأمر نفسه على بقية كلمات المحتوى مثل (قال، بن) مقارنة مع الكلمات الوظيفية المقاربة لها بعدد مرات التكرار.

ولو رجعنا إلى الجدول (٧) فسنجد أن حروف الجر بالذات، سواء كانت مفردة «في، من(٢٩)، على، عن..»، أو متصلة بضمير «عليه، له، فيها، بها..»، هي الأكثر تكراراً من بين بقية الأدوات(٣٠). وهذا التكرار العالي قد يعود إلى سببين، أحدهما تركيبي، والآخر وظيفي. أما التركيبي فهو أن حروف الجر من أهم وسائل تعدية الأفعال، كما أن شبه الجملة المتكون من الجار والمجرور من أكثر أساليب الاتساع في بناء الجملة استخداما. وأما الوظيفي فهو أن حروف الجر أساسية في أداء كثير من المعاني الوظيفية في اللغة. ومن بين حروف الجر، نجد أن حرف (في) بالذات هو أكثرها تكراراً. وقد يعود ذلك أيضاً لسببين. أولها أن (في) أساسية في أداء معنى الظرفية، وهو من أكثر المعاني تكرراً في اللغة، كما يدل عليه تكرار عدد كبير من الظروف في الجدول (٧) (٣١). والثاني ربما يعود إلى قوة (في) وتمكنها في الدلالة على كثير من المعاني والتبادل مع الحروف الأخرى. وهذه من المسائل التي تحتاج إلى مزيد من البحث والدراسة.

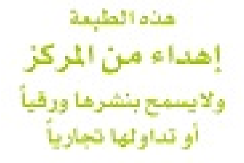
ومما يلفت الانتباه في الجدول أيضا أنه يحتوي على مصدرين اثنين فقط، هما «أيضا، خاصة». وكثرة تكرار هاتين الكلمتين عائدة إلى كثرة استخدامهما كأداتي ربط مما يجعلهما يقتربان كثيرا من الكلمات الوظيفية. وتحتوي القائمة كذلك على كلمة واحدة معربة، هي كلمة (مليون). وهي من أسماء الأعداد التي اقترضتها العربية في العصر الحديث، وأصبحت تستخدم بكثرة لعدم وجود كلمة بديلة في معناها. كما أصبحت تتصرف وفق النظام العربي، فتعرب

حسب موقعها في الجملة، وتثنى (مليونان/مليونين)، وتجمع جمع تكسير على (ملايين)، رغم أنها لا تجمع لأنها من قبيل غير المعدود في لغتها الأصلية.

وكذلك تحتوي القائمة على عدد من الظواهر والأساليب المرتبطة بـ«الفصحى المعاصرة» نتيجة للتأثر بالترجمة. فعلى سبيل المثال، يرد ضمن القائمة تعبير «بشكل»، وهو من الأساليب المستحدثة التي يكثر استخدامها بدلاً عن «الحال» في التركيب المعاصر. ويلفت الانتباه كذلك كثرة استخدام الفعل «تم». وكثرة استخدام هذا الفعل بالذات تعود إلى الاستعاضة به عن المبني للمجهول(٣٢)، كما يوضح الجدول (٩). واستخدام الظرف (حول)، كما يوضح الجدول (١٠)، متأثر أيضا بأساليب الترجمة المستحدثة لأن (حول) هنا ترجمة لكلمة (about). وكذلك الظرف (خلال) يكثر استخدامه في التعبير الحديث (من خلال)، ولذا فإن (من) وردت قبل (خلال) ١٩٣٩مرة، كما يوضح الجدول (١١).

العوضي طبيعة الصورة المفبركة والتي	تم	التلاعب فيها وتاريخها منذ عام
أن نشهد بأن هناك كاتبا	تم	ترهيبه لمجرد كتابة قارب فيها
تحدد القراءة عند عنوان جمالي	تم	فهمه بشكل يرضي غرور ذلك
الذهب فيها ومن جانب ثالث	تم	إدخال ورقة القيقب في كندا
مختلفا سنويا حتى عام حين	تم	تجميد التصميم ولكن بعد طلب
ومنها المسكوكة الفضية التذكارية التي	تم	اصدارها لتحمل ذكرى مناسبة مرور
أول المتاحف في الخليج وقد	تم	افتتاحه في عام حيث يعرض
البحرين وإلى جانب قاعاتها التخصصية	تم	إضافة المركز الإقليمي للتراث العالمي
المرممة وصل عددها إلى بيتا	تم	حمايتها من معول الهدم والسعي
على التقدير الذي يستحقونه وقد	تم	اختياره ضمن عشرة أفلام للتنافس

الجدول (٩) عينة من الكشاف السياقي لاستخدام كلمة (تم)



محور الكتابة المسرحية الجديدة المضامين	حول	ايضا تقديم أربع أوراق عمل
أثر العوامل الاقتصادية والاجتماعية والسياسية	حول	خصص جزؤها الأول لحديث نظري
الكيفية التي بدأ بها الكلام	حول	سنة ولا يوجد دليل واحد
ملابسات مقتل الطفلة وديمة من	حول	أن بدأت بعض المعلومات ترد
البنية السردية الروائية ونقاط التحول	حول	فصول بالعناوين التالية إضاءات عامة
الموضوع فالصفحات في معظمها اقتباسات	حول	خاليا من رؤية الباحث الشخصية
الأدب والنقد الأدبي والأسرة والمرأة	حول	الدورة الثانية عشرة دار الحديث
اللغة الكردية أنموذجا لمؤلفه آزاد	حول	كائن حي رؤية ونظرة فكرية
تراث وفكر مختلف الثقافات العالمية	حول	في عملية اقتناص بعض المعرفة
مجلة ما هو في نتاجها	حول	شعرية معينة التفاف الحركة الشعرية
المسرح العربي متخلف عن نتاجات	حول	المسرحي العربي الخطاب الذي يطرح

الجدول (١٠) عينة من الكشاف السياقي لاستخدام كلمة (حول)

ترتيب الكلمات السابقة قبل (خلال) حسب عدد مرات ورودها		ترتيب الكلمات اللاحقة بعد (من) حسب عدد مرات ورودها	
الكلمة	العدد	الكلمة	العدد
من	١٩٣٩	خلال	١٩٣٩
ومن	٩٨	أجل	٩٥١
دولار	٤٨	قبل	٩٠١
ذلك	٢٢	حيث	٤٠٧
جنيه	١٩	أن	٣٩٠
فمن	١٧	هذا	٣٨٧

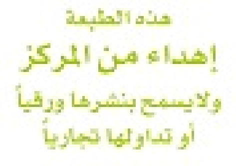
الجدول (١١) تصاحب (من) و(خلال) في المدونة

٥. ب. الكلمات الأكثر تكراراً في النصوص المتخصصة

كما رأينا في الجزء السابق، فإنه على الرغم من ورود عدد من كلمات المحتوى ضمن قائمة الكلمات الأكثر تكراراً إلا أن التكرار النسبي للنصوص التي وردت فيها قليل مقارنة بالكلمات الوظيفية. وهذا يعني أن كلمات المحتوى مرتبطة بنوعية معينة من النصوص تكثر فيها أكثر من غيرها. ولذا لا يكفي الكشف عن الكلمات الأكثر تكراراً بشكل عام، بل لا بد من الكشف عن الكلمات المتكررة في المجالات المتخصصة داخل المدونة، لمعرفة ما يرتبط ويكثر في تلك المجالات من كلمات. وتعرض الجداول (١٢)، (١٣)، (١٤)، (١٥)، (١٦)، (١٧) الكلمات الشائعة في النصوص المتخصصة في المجالات الستة التي وزعت عليها نصوص المدونة.

#	الكلمة									%
١٠	في	من	على	أن	إلى	التي	عن	ما	الذي	لا	٪١١,٨٧
٢٠	هذا	أو	هذه	مع	و	كان	بين	كل	كما	هو	٪٣,٢٥
٣٠	ذلك	لم	ان	العربية	الى	بعد	حيث	هي	العربي	خلال	٪١,٩٤
٤٠	وفي	كانت	الثقافة	بن	محمد	ولا	وهو	حتى	إن	فيها	٪١,٥٨
٥٠	له	الله	قد	ومن	عام	ثم	علي	وهي	فيه	بها	٪١,٣١
٦٠	إلا	الثقافي	عبد	وقد	أنه	بعض	الثقافية	به	الكتاب	العالم	٪١,١٥
٧٠	غير	اللغة	عليه	بل	تلك	قبل	الشاعر	اليوم	لها	يكون	٪١,٠١
٨٠	منها	أي	يمكن	الذين	لكن	م	أكثر	منذ	الحياة	العام	٪٠,٨٨
٩٠	هناك	العمل	وما	عبر	أما	الشعر	دون	فقد	أيضا	مثل	٪٠,٧٩
١٠٠	حول	المجتمع	عليها	بأن	يقول	أخرى	المسرح	كتاب	الشيخ	قال	٪٠,٧١
	مجموع التكرار النسبي										٪٢٤,٤٩

الجدول (١٢) الكلمات الأكثر تكراراً في النصوص الثقافية



#	الكلمة										٪
١٠	في	من	على	أن	الذي	التي	إلى	مع	عن	بعد	٪١١,٤٩
٢٠	و	ان	المنتخب	ما	الى	هذا	الفريق	لا	المباراة	كان	٪٣,١٤
٣٠	لم	هذه	حيث	كل	قبل	البطولة	خلال	فريق	اللاعبين	مباراة	٪٢,٢٠
٤٠	الاتحاد	وفي	الأول	بن	أمام	بين	اليوم	الثاني	القدم	ذلك	٪١,٧٠
٥٠	كما	أو	محمد	كرة	كأس	وهو	هو	الفوز	علي	حتى	٪١,٤٩
٦٠	الكرة	كانت	العالم	لكن	المركز	رئيس	له	منتخب	ثم	بطولة	٪١,٢٨
٧٠	نقطة	لكرة	الشوط	الدور	الثانية	العام	م	منتخبنا	الوطني	الموسم	٪١,١٢
٨٠	قد	المدرب	ولا	الأولى	عبد	المجموعة	اللاعب	النادي	الرياضية	منذ	٪١,٠١
٩٠	فى	وكان	أنه	وقال	اللقاء	ومن	الدوري	اللجنة	أي	يوم	٪٠,٩٤
١٠٠	نادي	عندما	الدقيقة	المشاركة	فيها	لاعب	فيما	بعض	عليه	عام	٪٠,٨٩
مجموع التكرار النسبي											٪٢٥,٢٧

الجدول (١٣) الكلمات الأكثر تكراراً في النصوص الرياضية

#	الكلمة										٪
١٠	في	من	على	أن	إلى	التي	عن	خلال	مع	ما	٪١٢,٣٩
٢٠	هذا	هذه	الى	الذي	العام	ان	مليون	بين	لا	حيث	٪٣,١٠
٣٠	و	كما	مليار	دولار	أو	القطاع	المالية	إن	ذلك	بنسبة	٪٢,٠٤
٤٠	بعد	عام	الحكومة	غير	الاقتصاد	السوق	كل	العمل	الاقتصادية	وقال	٪١,٥٠
٥٠	نحو	لم	الماضي	وفي	شركة	الشركات	تم	قبل	أي	قد	٪١,٢٩
٦٠	الاقتصادي	الدول	عدد	أنه	هو	أسعار	بعض	بشكل	مجلس	بلغت	٪١,١٦
٧٠	العامة	كان	قطاع	ألف	رئيس	الشركة	وهو	خاصة	هناك	دول	٪١,٠٥
٨٠	أكثر	النفط	البنك	نسبة	حتى	الدولة	وذلك	هي	علي	وزارة	٪٠,٩٩
٩٠	والتي	منها	الفترة	الأول	الوطني	الخاص	العالمية	منذ	ريال	مستوى	٪٠,٩٢
١٠٠	جديدة	المائة	زيادة	كانت	الاستثمار	سعر	دينار	عبر	العربية	مقارنة	٪٠,٨٦
مجموع التكرار النسبي											٪٢٥,٣٠

الجدول (١٤) الكلمات الأكثر تكراراً في النصوص الاقتصادية

#	الكلمة										%
١٠	في	من	على	أن	إلى	التي	عن	ما	لا	الذي	١٢,٠٩٪
٢٠	مع	ان	هذا	هذه	أو	الى	كان	ذلك	لم	بين	٣,٣٥٪
٣٠	الرئيس	كل	بعد	إن	هو	كما	النظام	رئيس	الشعب	وقال	٢,٢٢٪
٤٠	خلال	أي	كانت	قبل	و	الحكومة	وفي	مجلس	السوري	قد	١,٦٤٪
٥٠	المجلس	مصر	السياسية	حتى	العام	الدولة	هي	أنه	هناك	غير	١,٤١٪
٦٠	ولا	سوريا	علي	وهو	العربية	لكن	له	السورية	الثورة	المتحدة	١,٢٢٪
٧٠	حيث	بعض	السياسي	منذ	عليه	العربي	يمكن	مرسي	ومن	الأمن	١,١١٪
٨٠	الجيش	الوطني	اليوم	فيها	حول	يكون	الله	تلك	السلطة	محمد	١,٠٠٪
٩٠	تم	السابق	أكثر	الانتخابات	الآن	بأن	لها	وقد	فيه	وأن	٠,٩٣٪
١٠٠	الذين	قال	ضد	بها	دولة	إلا	فإن	عام	دون	المعارضة	٠,٨٧٪
	مجموع التكرار النسبي										٢٥,٨٣٪

الجدول (١٥) الكلمات الأكثر تكراراً في النصوص السياسية

#	الكلمة										%
١٠	في	من	الله	على	أن	لا	إلى	ما	عليه	عن	١٣,٢٥٪
٢٠	أو	التي	قال	ولا	هذا	الذي	صلى	وسلم	ذلك	كان	٣,٩١٪
٣٠	هذه	إن	كل	هو	الناس	له	تعالى	إلا	بن	بين	٢,٨٥٪
٤٠	ومن	به	مع	لم	كما	وهو	حتى	الإسلام	فيه	رسول	٢,١٧٪
٥٠	ثم	إذا	و	الذين	فيها	القرآن	وقد	أنه	وفي	بعد	١,٦٧٪
٦٠	بها	المسلمين	هي	وما	قد	يكون	النبي	الدين	عنه	بعض	١,٤٤٪
٧٠	محمد	يا	فإن	ابن	الإنسان	الى	يقول	يوم	عند	بل	١,٢٧٪
٨٠	ان	غير	أي	فقال	الإسلامية	وقال	لهم	فلا	فقد	الإسلامي	١,١٥٪
٩٠	الصلاة	الكريم	لها	رمضان	وأن	عبد	كانت	علي	وهي	حيث	١,٠١٪
١٠٠	الرسول	لأن	رضي	الشيخ	ولكن	العلم	منها	أهل	بما	نفسه	٠,٩٢٪
	مجموع التكرار النسبي										٢٩,٦٤٪

الجدول (١٦) الكلمات الأكثر تكراراً في النصوص الدينية

#	الكلمة										٪
١٠	في	من	على	أن	إلى	التي	عن	أو	ما	هذه	١٢,٦٥٪
٢٠	هذا	مع	لا	الذي	و	كما	بين	خلال	حيث	قد	٣,٠٢٪
٣٠	ذلك	بعد	يمكن	كل	ان	هو	أكثر	الى	مثل	إن	١,٩٠٪
٤٠	غير	كان	بعض	بشكل	أي	وهو	العالم	وفي	الدم	عند	١,٤٥٪
٥٠	وقد	هي	ومن	أنه	قبل	حتى	لم	يكون	كانت	وهي	١,٢٥٪
٦٠	أيضا	استخدام	عام	هناك	الجسم	أخرى	تكون	علي	الطاقة	الذين	١,٠٨٪
٧٠	الدراسة	تلك	يتم	بها	إذا	ولكن	العام	تم	فإن	الصحية	٠,٩٨٪
٨٠	عبر	وذلك	منها	شركة	عدد	الدكتور	الجهاز	الإصابة	حول	لدى	٠,٩٢٪
٩٠	تناول	إلا	العلاج	المعلومات	مما	لها	الأطفال	أما	فيها	عملية	٠,٨٦٪
١٠٠	والتي	طريق	له	ثم	نسبة	دون	يجب	خاصة	لكن	المرض	٠,٧٩٪
مجموع التكرار النسبي											٢٤,٩١٪

الجدول (١٧) الكلمات الأكثر تكراراً في النصوص العلمية

أول ملاحظة تبدو للناظر في الجداول السابقة هي أن كلمات المحتوى بدأت تظهر بشكل أكبر ضمن قوائم الكلمات الأكثر تكراراً. وهذا متوقع، ففرز النصوص وتوزيعها بحسب المجالات التي تنتمي إليها يؤدي إلى ظهور كلمات المحتوى التي تعبر عن تلك الموضوعات بشكل أكبر. ولكن يلاحظ أيضا أن أغلب كلمات المحتوى المتكررة من الأسماء، أما الأفعال فلا تزال قليلة جداً في جميع الجداول. وهذه الأسماء منها ما هو مشترك بين اثنين أو أكثر من الجداول السابقة، ومنها ما لم يرد إلا في جدول واحد. وسنترك حاليا الأسماء التي وردت متكررة في أكثر من مجال، لأننا سنعلق على بعضها في ثنايا الكلام لاحقا. وأما الأسماء التي لم ترد ضمن الكلمات الأكثر تكرارا إلا في جدول واحد فقط، فهي كثيرة ومتنوعة. ويقدم الجدول (١٨) تلخيصا بتلك الأسماء مع توزيعها على المجالات التي وردت فيها.

المجال	الكلمات التي انفرد المجال بظهورها ضمن الأكثر تكراراً
النصوص الثقافية	المجتمع، الحياة، الثقافة، الثقافي، الثقافية، الكتاب، كتاب، اللغة، الشاعر، الشعر، المسرح
النصوص الرياضية	المنتخب، منتخب، منتخبنا، الفريق، فريق، الاتحاد، المباراة، المجموعة، البطولة، بطولة، اللاعبين، اللاعب، لاعب، القدم، الكرة، كرة، لكرة، كأس، الفوز، المركز، نقطة، الشوط، الدور، الموسم، المدرب، النادي، نادي، الرياضية، اللقاء، الدوري، اللجنة، الدقيقة، الثانية، المشاركة
النصوص الاقتصادية	الدول، دول، الماضي، العامة، دولار، ريال، دينار، مليون، مليار، ألف، المائة، شركة، الشركة، الشركات، المالية، وزارة، القطاع، قطاع، بنسبة، الاقتصاد، الاقتصادية، الاقتصادي، السوق، أسعار، سعر، النفط، البنك، فترة، الخاص، مستوى، زيادة، الاستثمار، مقارنة
النصوص السياسية	الرئيس، دولة، المجلس، مصر، النظام، الشعب، السياسة، السياسي، المتحدة، السوري، سوريا، السورية، الثورة، مرسي، الأمن، الجيش، السلطة، الانتخابات، ضد، المعارضة، السابق
النصوص الدينية	ابن، الناس، الإنسان، الدين، الإسلام، المسلمين، الإسلامي، رسول، الرسول، القرآن، النبي، الصلاة، رمضان، العلم، أهل، نفسه، الكريم
النصوص العلمية	الدكتور، الدم، استخدام، الجسم، الطاقة، الدراسة، الصحية، الجهاز، الإصابة، العلاج، المعلومات، الأطفال، عملية، طريق، المرض

الجدول (١٨) الأسماء الأكثر تكراراً في المجالات المختلفة

وكما يظهر في الجدول السابق، يوجد نوعية معينة من الكلمات المتكررة في كل مجال تعكس طبيعة الموضوعات المطروقة فيه. فيلاحظ -على سبيل المثال- أن الكلمات الأكثر تكراراً في النصوص العلمية ذات ارتباط بالجوانب الطبية والصحية. وللكلمات الأكثر تكراراً في النصوص السياسية ارتباط كبير بالأحداث الجارية. فكلمات (مصر، السوري، سوريا، السورية، مرسي، السابق) ربما لم تتكرر إلا بسبب طبيعة الأحداث التي تزامنت مع تاريخ نصوص المدونة، بخلاف الكلمات في المجالات الأخرى التي يبدو أن لها ارتباطاً أكثر دواما بالمجالات التي ظهرت فيها. ومن الناحية الكمية، يلاحظ أن النصوص الرياضية والاقتصادية بالذات تحتوي على عدد أكبر من الأسماء ضمن قائمة الكلمات الأكثر تكراراً. وربما يعود ذلك إلى أن هذين المجالين مجالان متخصصان لهما كلمات خاصة محصورة ومحدودة مما يجعلها تتكرر كثيراً. وسنجد ما يؤكد هذه النقطة بالنسبة للمجال الرياضي بشكل خاص في الجزء التالي المتعلق بالكلمات المميزة للمجالات المختلفة.

٥. ج. الكلمات المميزة للمجالات المختلفة

لا تقتصر إمكانيات البحث في المدونات على الكلمات الأكثر تكراراً في المجالات المختلفة، بل يمكن أيضا استخراج الكلمات المميزة لكل مجال من المجالات على حدة. واستخراج الكلمات المميزة يتطلب أن يكون لدينا مدونتان، أحدهما هي المدونة التي نريد معرفة ما يميزها من كلمات والأخرى هي المدونة المرجعية، وتحوي النصوص التي تستخدم للمقارنة مع نصوص المدونة التي نريد تمييزها. ويتم استخراج الكلمات المميزة عن طريق مقارنة التوزيع الإحصائي للكلمات في المدونة التي نريد تمييزها بالكلمات في المدونة المرجعية.

وفي دراستنا الحالية قمنا باستخراج الكلمات المميزة عن طريق مقارنة كل مجال من المجالات الستة ببقية نصوص المدونة الكاملة التي تشمل المجالات

الأخرى ما عدا المجال المقصود. ومع أنه يوجد طرق إحصائية متعددة للمقارنة واستخراج الكلمات المميزة، إلا أننا استخدمنا طريقة «معامل الغرابة» لأنها من أسهل الطرق وأوضحها. ومعامل الغرابة هو حاصل قسمة التكرار النسبي للكلمة في المدونة التي نريد تمييزها (المدونة أ) على التكرار النسبي للكلمة نفسها في المدونة المرجعية (المدونة ب). وحاصل القسمة قد يكون:

أ - (ما لا نهاية): إذا ظهرت الكلمة في المدونة (أ) فقط ولم تظهر في المدونة (ب).

ب -أو (رقماً كبيراً جداً): إذا كان التكرار النسبي للكلمة في المدونة (أ) أكبر بكثير من تكرارها النسبي في المدونة (ب)

ج - أو (١): إذا تساوى التكرار النسبي للكلمة في المدونتين

د - أو (قريباً من الصفر): إذا كان التكرار النسبي للكلمة في المدونة (ب) أكبر بكثير من تكرارها النسبي في المدونة (أ).

ويتم اعتبار الكلمة من الكلمات المميزة عندما تكون قيمة معامل الغرابة (ما لا نهاية)، أو أكبر من رقم معين يتم اختياره ووضعه كمقياس للمقارنة. وفي حالتنا فقد تم اختيار الرقم ١٠ كحد أدنى لقيمة معامل الغرابة المقبول للكلمات المميزة.

ولكن لا بد من ملاحظة أنه من الممكن أن تكون قيمة معامل الغرابة (مالا نهاية) حتى ولو لم تظهر الكلمة إلا لمرة واحدة فقط في المدونة (أ) ولم تظهر في المدونة (ب). ولتجنب ذلك لابد من وضع معيار آخر غير قيمة معامل الغرابة يتعلق بعدد مرات تكرار الكلمة من أجل إدراجها ضمن الكلمات المميزة. وفي حالتنا فقد اخترنا الرقم (٤٠) ليكون الحد الأدنى لعدد مرات التكرار المطلوبة لتضمين الكلمة في قائمة الكلمات المميزة.

وبناء على هذين المعيارين فقد تم استخراج الكلمات المميزة في كل مجال على حدة. كما تم توزيع الكلمات المميزة على أربعة تدرجات أو مستويات، كما توضح الجداول (١٩)، (٢٠)، (٢١)، (٢٢)، (٢٣)، (٢٤). فالمستوى الأول يشمل الكلمات التي قيمة معامل غرابتها هي (ما لا نهاية)، وقد كتبت في الجداول بالخط الغامق. وأما المستوى الثاني فيشمل الكلمات التي معامل غرابتها أكبر من أو يساوي (٥٠) وليس ما لا نهاية، وكتبت بالخط المائل. والمستوى الثالث يشمل التي معامل غرابتها أكبر من أو يساوي (١٠) وأقل من (٥٠)، كتبت وتحتها خط. وأما المستوى الأخير، فيشمل الكلمات التي معامل غرابتها أقل من (١٠)، وقد كتبت بدون تمييز، ولا تعد من قبيل الكلمات المميزة، لأنها لم تتجاوز الحد الأدنى لمعامل الغرابة المطلوب.

القصيدة	**المساري**	**التشكيلية**	**المخطوطات**	**للمسرح**	**زيزيك**	*المسرحية*	*المثقف*	*المسرح*	*للشاعر*
الإبداعي	ودان	*الموسيقي*	*الشعرية*	*الشعري*	شنقيط	*للفنون*	*السرد*	قصيدة	*المسرحي*
الروائي	*الفنان*	*والشاعر*	*المثقفين*	*الفنانين*	*القصائد*	قصائد	شعرية	*الشاعر*	*السينمائي*
الناقد	*الفنون*	*التشكيلي*	*الأدباء*	*الصالون*	*الدراما*	شعراء	*السينما*	مسرح	*الأدبي*
الثقافي	*الثقافة*	*الأديب*	*ثقافي*	المبدعين	الجابري	المكتبة	الرواية	الفيلم	الكتابة
الأدبية	مسرحية	الإبداعية	المكتبات	للكتاب	النسوية	ثقافية	المتلقي	الإبداع	للثقافة
والتراث	والفنون	الفن	الثقافية	الشعراء	درويش	الحداثة	الأدب	اللغوي	الكاتب
المبدع	الفلسفة	النقاد	شاعر	اللغوية	الشعر	التراث	والفن	الموسيقية	أدب
الأفلام	مكتبة	اللغات	فن	المخرج	اللغة	القارئ	الثقافات	ديوان	الحب
فرقة	المختار	الكتب	نوبل	الترجمة	والثقافة	الجائزة	مهرجان	التدريس	رواية

الجدول (١٩) الكلمات المميزة لنصوص المجال الثقافي في المدونة

الشوط	**منتخبنا**	**المنتخبات**	**التأهل**	**للبطولة**	**اللاعبون**	**لاعبو**	**الاولمبي**	**العنابي**	**ميدالية**
لاعبيه	**يونايتد**	**رونالدو**	**تأهل**	**بالتعادل**	**منتخبات**	**الرجوب**	**باللقب**	**مضيفه**	**ميداليات**
مباراته	**إب**	**الهجومية**	**سدد**	**ركلة**	**تمريرة**	**اولمبياد**	**بالبطولة**	**الكروية**	**لدوري**
برونزية	**لمنتخبنا**	**ملعبه**	**لاعبينا**	**الترجي**	**بولندا**	**ركنية**	**هداف**	**الهجومي**	**بوسكي**
للناشئين	**للمنتخبات**	**سددها**	*لاعبين*	*المباراة*	*الأولمبي*	*الاندية*	*التصفيات*	*الملاعب*	*بهدفين*
برصيد	*مانشستر*	*البطولة*	*للمنتخب*	مهاجم	*للكرة*	*المدربين*	*المدرب*	*الكروي*	*البطولات*
المنافس	*الكابتن*	*المرمى*	منافسات	مدرب	*للمباراة*	رصيده	*الفيفا*	*الفريقان*	*الأولمبياد*
اللاعبين	*ارسنال*	*التعادل*	*الزمالك*	*الاتحادات*	فريقي	*المونديال*	*لكرة*	*لاعبي*	*الوداد*
مباراة	مباريات	*المهاجم*	ريكاردو	بطولة	*الميداليات*	*الحارس*	برشلونة	*لبطولة*	*الأسباني*
الآسيوي	بفوزه	ميسي	*البرازيلي*	*اللقب*	*الودية*	تصفيات	*لكأس*	فريقه	*اللاعب*

الجدول (٢٠) الكلمات المميزة لنصوص المجال الرياضي في المدونة

الشوط	منتخبنا	المنتخبات	التأهل	للبطولة	اللاعبون	لاعبو	الاولمبي	العنابي	ميدالية
لاعبيه	يونايتد	رونالدو	تأهل	بالتعادل	منتخبات	الرجوب	باللقب	مضيفه	ميداليات
مباراته	إب	الهجومية	سدد	ركلة	تمريرة	اولمبياد	بالبطولة	الكروية	لدوري
برونزية	لمنتخبنا	ملعبه	لاعبينا	الترجي	بولندا	ركنية	هداف	الهجومي	بوسكي
للناشئين	للمنتخبات	سددها	لاعبين	المباراة	الأولمبي	الاندية	التصفيات	الملاعب	بهدفين
برصيد	مانشستر	البطولة	للمنتخب	مهاجم	للكرة	المدربين	المدرب	الكروي	البطولات
المنافس	الكابتن	المرمى	منافسات	مدرب	للمباراة	رصيده	الفيفا	الفريقان	الأولمبياد
اللاعبين	ارسنال	التعادل	الزمالك	الاتحادات	فريقي	المونديال	لكرة	لاعبي	الوداد
مباراة	مباريات	المهاجم	ريكاردو	بطولة	الميداليات	الحارس	برشلونة	لبطولة	الأسباني
الآسيوي	بفوزه	ميسي	البرازيلي	اللقب	الودية	تصفيات	لكأس	فريقه	اللاعب

الجدول (٢١) الكلمات المميزة لنصوص المجال الاقتصادي في المدونة

أنان	اشتباكات	العدلي	جنيف	نتنياهو	هجليج	الجنائية	المخابرات	كلينتون	الدستوري
الاسد	كوفي	المدنيين	اسرائيل	الإرهابية	بوتين	للجيش	الدستورية	دارفور	الجمهوري
اغتيال	الأسد	قوات	مرسي	كردفان	القوات	للقوات	المسلحة	الاقتراع	للانتخابات
عسكري	القذافي	عرفات	للحزب	الإخوان	شفيق	الرئاسي	الرئاسية	الانتقالية	عبدربه
المراقبين	بشار	عنان	عسكرية	هادي	السورية	المعارضة	انتقالية	التأسيسية	السوري
لحزب	الانتخابات	جرائم	أحزاب	حمص	العسكري	بعودة	المرزوقي	المحكمة	الامن
أفغانستان	العسكرية	الأغلبية	الإسرائيلي	ليبيا	البرلمان	الاخوان	السوريين	القضائية	الجيش
طهران	انتخابات	سوريا	الاحتجاجات	حزب	قيادات	مقتل	الدستور	التوافق	الانتخابية
موسكو	المستشار	النائب	لحقوق	نواب	التيار	للسلطة	الديمقراطي	التدخل	الإيرانية
الحزب	البرلمانية	ونقلت	الجرائم	الجماعة	الإصلاحات	إسرائيل	جريمة	سورية	للشعب

الجدول (٢٢) الكلمات المميزة لنصوص المجال السياسي في المدونة

هريرة	عباده	الترمذي	الصدقة	الرسل	رواه	البقرة	التوبة	الذنوب	رسول
القيامة	الفتاوى	صلى	إله	المشركين	الزكاة	البخاري	وسلم	المؤمنين	أمرنا
تيمية	رضي	الصحابة	تعالي	أخرجه	الآخرة	صدقة	النبي	ربك	الآية
والصلاة	الدعاء	آمنوا	الصلاة	عليم	ورسوله	سورة	زكاة	للمسلمين	تعالى
العبادات	ربنا	ربهم	بالمعروف	سبحانه	بسم	وآله	المؤمن	وقوله	فلما
المنكر	وجل	الآيات	الرسول	العبادة	الإيمان	والله	تبارك	الجنة	قريش
عبادة	أنزل	اللهم	والسلام	بالله	لقوله	الشرع	آيات	قلوبهم	الإمام
أجر	منكم	وتعالى	الأئمة	المذاهب	مسلم	العالمين	النبوي	المساجد	والنهي
آية	الشيطان	أنس	أخلاق	الخلق	خيرا	رب	الحرام	بالدين	نبي
رحمه	الله	الفقهية	القلوب	القرآن	صلاة	كنتم	المسلم	الجهاد	الأنبياء

الجدول (٢٣) الكلمات المميزة لنصوص المجال الديني في المدونة

أندرويد	الغدة	آبل	الجراحة	فيتامين	الروبوتات	الروبوت	الحواسيب	ويندوز	الجينات
نوكيا	الجينية	الدرقية	إنتل	الأسبرين	التأكسد	الهضمي	الجذعية	غوغل	الصداع
جراحة	الدهون	القولون	فيروس	حاسوب	هرمون	السمنة	أبل	السكري	الأورام
العدسات	السوائل	الاكتئاب	الاصطناعي	السرطانية	العلاجات	جوجل	القلبية	الكلوي	الثدي
التهاب	مضاعفات	فيسبوك	الكالسيوم	الأعشاب	المعالج	اللوحية	البرمجيات	هاتف	الألياف
سرطان	المعدية	الأشعة	الكبد	الكلى	الهضم	الشرايين	مستخدم	المستخدم	المناعة
الفيروس	الباحثون	بسرطان	الآلام	بمرض	بأمراض	المضادات	الفيروسات	الأدوية	بيري
البول	للجسم	البكتيريا	مايكروسوفت	الخلايا	الوراثية	الهواتف	المريض	طب	مرضى
الأعراض	التدخين	الأسنان	الكمبيوتر	باحثون	العصبي	مستخدمي	السرطان	مريض	الذكي
المستخدمين	أمراض	المحمولة	الحاسوب	الدم	الدموية	أدوية	الجسم	الحرارية	لعلاج

الجدول (٢٤) الكلمات المميزة لنصوص المجال العلمي في المدونة

أول ملاحظة تبدو لنا هي أن توزيع الكلمات المميزة بالنسبة لمستويات معامل الغرابة مختلف بين المجالات، كما يلخص الجدول رقم (٢٥). وكما يبدو فإن أكثر المجالات تميزاً هو المجال الرياضي حيث تحتوي النصوص التي تنتمي إلى هذا المجال على عدد كبير نسبيا من الكلمات الخاصة (٤٣ كلمة) التي لم ترد أبداً في أي مجال آخر. كما أن بقية الكلمات المميزة في هذا المجال لها معامل غرابة كبير نسبيا. وهذا يؤكد الملاحظة التي ذكرناها سابقا، فالمجال الرياضي مجال مستقل بشكل واضح، له كلمات محصورة ومحددة تتكرر بشكل كبير، وتميزه إلى حد كبير عن غيره من المجالات الأخرى.

المدونة	معامل الغرابة			
	ما لا نهاية	أكبر من ٥٠	أكبر من ١٠	أقل من ١٠
الثقافية	٦	٣٨	٥٦	٠
الرياضية	٤٣	٥٧	٠	٠
الاقتصادية	٥	٤٣	٥٢	٠
السياسية	٢	١١	٨٣	٤
الدينية	٥	٣٩	٥٦	٠
العلمية	١٨	٧٣	٩	٠

الجدول (٢٥) عدد الكلمات المميزة بالنسبة لمعامل الغرابة في المجالات المختلفة

ويأتي بعد المجال الرياضي المجال العلمي، حيث تحتوي نصوص هذا المجال على (١٨) كلمة مميزة لم ترد في أي مجال آخر. كما أن كثيراً من كلماته ذات معامل غرابة مرتفع نسبيا. ورغم أن ذلك يعني أن نصوص المجال العملي متميزة إلى حد ما من ناحية نوعية الكلمات المستخدمة، إلا أن تميزه لم يكن بالقدر المتوقع. ولعل السبب في ذلك يعود إلى أن نصوص المجال العلمي في الصحافة ذات صبغة عامة، موجهة إلى عموم القراء في المجتمع، وليست نصوصاً متخصصة، موجهة إلى قراء متخصصين.

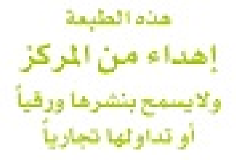
ومن ناحية أخرى، فإن المجال السياسي هو أقل المجالات تميزاً. فالكلمات المستخدمة في هذا المجال مشتركة مع بقية المجالات إلى حد كبير. وكما رأينا سابقا، فإن الكلمات الأكثر تكراراً في هذا المجال مرتبطة غالباً بالأحداث الجارية، وليس لها ارتباط دلالي دائم بالمجال السياسي. وربما يمكن أن نستنتج من ذلك أن المجال السياسي في الصحافة العربية هو أكثر المجالات عمومية، وأقلها نضجا بسبب افتقاره إلى الكلمات المحددة والخاصة، وهذه من المسائل التي تحتاج إلى مزيد من البحث والمقارنة.

ولو أردنا أن نقف على طبيعة الكلمات المميزة فسنجد أن أغلب الكلمات المميزة في المجال الثقافي لها دلالات ترتبط بالفنون والأنواع الأدبية المختلفة مثل الشعر والرواية والمسرح والرسم التشكيلي والسينما والموسيقا. وهناك مواد معينة تكثر تصريفاتها واشتقاقاتها ضمن الكلمات المميزة للمجال الثقافي (ثقف، أدب، بدع، فنن، كتب). وتحتوي الكلمات المميزة في هذا المجال على عدد من المعربات القديمة (ديوان، مهرجان، فلسفة، موسيقا) والحديثة (السينما، الدراما، الفيلم). وأما المجال الرياضي فكثير من كلماته المميزة عبارة عن مصطلحات موضوعة حديثا لدلالات مخصوصة في هذا المجال (الشوط، المنتخبات، التأهل، البطولة، المرمى، ركلة، تمريرة، ركنية). وكذلك تحتوي كلماته المميزة على

عدد من المعربات الحديثة(٣٣) ذات الدلالات المحصورة (المونديال، الأولمبياد، الفيفا، ميدالية، برونزية، الكابتن، الفيفا).

وأما المجال السياسي فنسبة كبيرة من كلماته المميزة عبارة عن أعلام، إما أسماء أشخاص أو دول أو أقاليم أو منسوبة إلى بعضها. ويوجد قليل من الكلمات المعربة في هذا المجال (الدستور، البرلمان، الديمقراطي). وكذلك المجال الاقتصادي لا يوجد ضمن كلماته المميزة إلا عدد قليل من الكلمات المعربة (بنك، بورصة) وأغلبها من قبيل أسماء العدد والعملات والمقادير (مليون، مليار، طن، برميل). بخلاف المجال العلمي الذي تطغى عليه بشكل واضح الكلمات المعربة والمقترضة. ورغم أن بعضها من قبيل الأعلام كأسماء الشركات والآلات (أندرويد، مايكروسوفت، آبل، غوغل، فيسبوك، ويندوز)، إلا أن كثيراً منها مصطلحات معربة (الروبوتات، الجينات، البرمجيات، الفيروسات، البكتيريا، السرطان، هرمون، التأكسد، الكالسيوم، القولون). وكذلك يوجد عدد كبير من المصطلحات الموضوعة حديثا (الهواتف، الحواسيب، الأشعة، الألياف، الخلايا، العدسات، المضادات، الجراحة). كما أن عدداً من الكلمات المميزة في المجال العلمي لها دلالات مرتبطة بالجسد وأعضائه ووظائفه (الجسم، الغدة، القلبية، الكلى، الثدي، الكبد، الشرايين، الأسنان، العصبي، الدم، الهضم، البول).

وكما يوضح الاستعراض السابق، فإن الغالبية العظمى من الكلمات المميزة في كافة المجالات من كلمات المحتوى التي تحمل غالبا دلالات حديثة. وأكثرها من الأسماء، وأما الأفعال فمحدودة جدا. وتحتوي الأسماء على عدد كبير من المصادر والمشتقات المولدة والمستحدثة. فمنها أبنية مستخدمة اكتسبت دلالات جديدة مثل (جراحة، المرمى، المثقف)، ولكن منها أيضا ما هو جديد في مبناه ومعناه مثل (سيولة، مسرح، محكمة، مهاجم، مخرج). فالمصدر (سيولة) – على سبيل المثال – بناء مستحدث على وزن فعولة في هذه المادة. وكذلك اسم



الفاعل (مهاجم) وفعله (هاجم) كلاهما بناء جديد مستحدث ومولد من مادة (ه ج م)، وينطبق ذلك على اسم المكان (محكمة) في مادة (ح ك م).

ورغم وجود نسبة كبيرة من الكلمات المعربة والمقترضة، إلا أنه يلاحظ وجود حالات قليلة وردت فيها الكلمات الأجنبية ومقابلاتها العربية المولدة معا ضمن الكلمات المميزة، مثل (بنك/مصرف) في المجال الاقتصادي، و(حاسوب/ كمبيوتر) في المجال العلمي. وكما يوضح الجدول (٢٦) فإن كلمتي (حاسوب) و(كمبيوتر) تستخدمان بمقدار متساو تقريبا. أما كلمة (بنك) فتستخدم بشكل أكبر بكثير من مقابلها العربي (مصرف)، خاصة حينما تكون معرف بالألف واللام. وقد يكون ذلك عائدا إلى أن كلمة (البنك) تستخدم كجزء من أسماء البنوك، مثل (البنك الأهلي، البنك الوطني...). ولذا لم ينجح المقابل العربي لهذه المفردة بسبب أن المؤسسات المقصودة تسمي نفسها بالمقابل الأجنبي.

نوع الكلمة	المصرف والبنك				الحاسوب والكمبيوتر			
	الكلمة	العدد	الكلمة	العدد	الكلمة	العدد	الكلمة	العدد
منكرة	مصرف	١٠٦	بنك	٢٦٩	حاسوب	٦٢	كمبيوتر	٢٩
معرفة	المصرف	٩٥	البنك	٥٢٩	الحاسوب	٩٨	الكمبيوتر	١٣٣
المجموع		٢٠١		٧٩٨		١٠٦		١٦٢

الجدول (٢٦) تكرار كلمتي (بنك وكمبيوتر) ومقابليهما العربيين في المدونة

٥. د. البحث في مادة معجمية: مادة (عمل) مثالاً

بخلاف ما سبق مما يعتمد على البحث في مجموع نصوص المدونة لاستكشاف واستخراج الكلمات الشائعة أو المميزة، فإن لسانيات المدونات يمكن أن تستخدم للبحث في قضية محددة سلفا كالبحث عن كلمة معينة، أو تعبير محدد، أو مادة مقصودة، لاستكشاف تصريفاتها واشتقاقاتها، ومواضع استخدامها، وما يطرأ عليها تبعا لذلك من اختلاف في المعاني أو طرق الاستخدام. وغالبا ما

يتم استخدام هذا النوع من البحث بناء على فرضيات معينة متعلقة بالكلمات والعبارات المقصودة. فلو أخذنا مادة (ع م ل)، على سبيل المثال، فسنجد أن كلمة (العمل) قد وردت ضمن الكلمات الأكثر شيوعا في المدونة كاملة (جدول ٧)، وفي مجالين من المجالات الخاصة، وهما المجالان الثقافي والاقتصادي (الجدولان ١٢ و ١٤). وقد ورد من مشتقات هذه المادة كلمة (عملية) ضمن الكلمات الأكثر شيوعا في المجال العلمي (جدول ١٧)، وكلمة (تعاملات) ضمن الكلمات المميزة للمجال الاقتصادي (جدول ٢١).

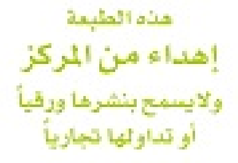
ورغم أن مادة (ع م ل) مادة ثرية ومتشعبة في المعجمات العربية، إلا أنه قد استجد فيها العديد من الأبنية والدلالات(٣٤). فمن الأبنية المستحدثة في هذه المادة المصدر الصناعي (عملية)، والصفة على وزن (فعيل) (عميل)، والمصدر على وزن (فعولة) (عمولة)، واسم المكان (معمل). وبالرجوع إلى المدونة يمكننا استخراج هذه الكلمات وتكرارها ومواضع ورودها. ورغم أن متابعة التحليل التفصيلي قد تؤدي إلى الكثير من النتائج، إلا أننا سنكتفي هنا فقط باستعراض تكرار هذه الكلمات في المدونة كاملة، وفي المجالات الفرعية المختلفة، كما يوضح الجدول (٢٧).

الكلمة	تصريفاتها	ثقف(٣٥)	ريض	قصد	سيس	دين	علم	كمل	المجموع
عملية	عملية	١٢٤	٦٢	٢٠٦	٢٢٤	٧٠	٢٨٧	٩٧٣	١٩٧٨
	العملية	٦٥	٢٠	٥٧	١٣٤	٢٧	٩٠	٣٩٣	
	عمليات	٣٢	١٣	١٨٤	١١١	٩	٨٧	٤٣٦	
	العمليات	١٤	٧	٣٣	٥٣	٩	٦٠	١٧٦	
عميل	عميل	١	٠	٣	٢	١	٧	١٤	١٧٩
	العميل	٠	٠	١٥	٤	٠	١٢	٣١	
	عملاء	٠	١	٢٢	٧	٠	١٠	٤٠	

	العملاء	٠	٠	٦٧	٧	٠	٢٠	٩٤	
معمل	معمل	١	١	١٠	٠	٠	٣	١٥	٥٢
	المعمل	٠	٠	٢	٠	٢	٥	٩	
	معامل	٢	٠	٧	٢	٠	٦	١٨	
	المعامل	١	٠	٣	١	١	٤	١٠	
عمولة	عمولة	٠	٠	٢	٠	٠	٠	٢	١٠
	العمولة	٠	٠	٠	٠	٠	٠	٠	
	عمولات	٠	٠	٢	٠	٠	٠	٢	
	العمولات	٠	٠	٦	٠	٠	٠	٦	
المجموع	٢٤٠	١٠٤	٦١٩	٥٤٥	١١٩	٥٩١	٢١١٩		

الجدول (٢٧) الأبنية الجديدة في مادة (عمل) وورودها في المدونة

وكما يلاحظ في الجدول السابق، فإن كلمة (عملية) هي الأكثر استخداما من بين الأبنية الجديدة في مادة (ع م ل)، وبفارق كبير جدا عن الأبنية الأخرى. وكلمة (عملية) من الكلمات المتأثرة بالترجمة فهي تستخدم في مقابل كلمتي (process) و (operation). وربما تعود كثرة استخدامها إلى أنها ذات مدى دلالي واسع ولا تختص بمجال معين. وأما بقية الكلمات فمحدودة الاستخدام جدا. فـ(عمولة) (٣٦)لم ترد إلا في عشرة مواضع فقط، ومحصورة في المجال الاقتصادي. وكذلك (عميل) ذات ارتباط بالمجال الاقتصادي للدلالة من يتعامل مع غيره في حرفة أو صنعة أو تجارة، وهي الأكثر استخداما في هذا المعنى في مقابل الكلمة التقليدية (حريف) و (زبون) وهي كلمة مولدة أخرى كما يوضح الجدول (٢٨). وأما (حريف) فرغم أنها الكلمة التراثية المستخدمة في معنى (عميل) (٣٧)، إلا أنها اندثرت تقريبا في الاستخدام المعاصر، إلا في مواضع محدودة جدا في المغرب العربي. فلم ترد في المدونة بهذا المعنى إلا في ثلاثة مواضع فقط يعرضها الجدول (٢٩).



الكلمة	تصريفاتها	التكرار	المجموع
عميل	عميل	١٤	١٧٩
	العميل	٣١	
	عملاء	٤٠	
	العملاء	٩٤	
زبون	زبون	٢	٧٨
	الزبون	١٢	
	زبائن	١٨	
	الزبائن	٤٦	
حريف	حريف	٠	٣
	الحريف	٢	
	حرفاء	٠	
	الحرفاء	١	

الجدول (٢٨) عميل ومرادفاتها في المدونة

الأوروبي تفيد أن تونس تعد	الحريف	الثالث والعشرين للمغرب وبينما تحتل
أما حصة المغرب التي تعد	الحريف	الحادي عشر لتونس والسابع والثلاثين
اراء المضيفين والمضيفات وكذلك بعض	الحرفاء	الاوفياء المتعلقة باقتراحات موجهة للادارة

الجدول (٢٩): المواضع التي ودت فيها (حريف) بمعنى (عميل)

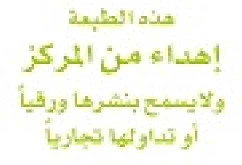
وفي مقابل هذه الأبنية الجديدة، هناك كلمات أخرى في مادة (ع م ل) ذات أبنية كانت مستخدمة سابقا ولكن بدلالات مختلفة عن دلالاتها المعاصرة. ومن ذلك كلمتا (عُمّلة) و(عمالة) «مثلثة الفاء»، فهاتان الكلمتان كانتا تعنيان في السابق «أجرة العامل على ما بذله من عمل». ولكن تحول معنى (عملة) وأصبحت تستخدم في الدلالة على نوعية النقد المستعمل في التبادل التجاري والاقتصادي،

في مقابل كلمة (currency) الإنجليزية. وكذلك (عمالة) أصبحت تدل غالبا على مجموع القوى العاملة، ويلخص الجدول (٣٠) تكرار هاتين الكلمتين في المدونة ومجالاتها الفرعية.

الكلمة		ثقف	ريض	قصد	سيس	دين	علم	كمل
عملة	عملة	٩	٠	١٢	٥	٣	٢	٣١
	العملة	١٤	٢	٨٧	٨	٣	٠	١١٤
عمالة	عمالة	١	٠	١١	٤	٠	٠	١٦
	العمالة	٥	٠	٤٢	١٥	٢	٣	٦٧
المجموع		٢٩	٢	١٥٢	٣٢	٨	٥	٢٢٨

الجدول (٣٠) تكرار كلمتي (عملة) و(عمالة) في المدونة

ومن ناحية أخرى، فقد لا تتحول دلالة المفردة بشكل كامل، ولكن يلحقها نوع من التقييد والتخصيص. وعلى سبيل المثال، فاسم الفاعل لوصف العاقل من (ع م ل) (عامل) الذي يدل على من يقوم بالعمل أو يشتغل في حرفة معينة له جمعا تكسير، هما: عَمَلة، وعُمّال. وفي القديم، كانت كلمة (عُمّال) تستخدم للإشارة إلى الولاة ومن يلون أعمال السلطان، مثل (عمال الصدقة، وعمال الأمصار)، وأما (عَمَلة) فكان يدل غالبا على من يشتغلون بأيديهم في أعمال تتطلب الجهد البدني كالحفر ونحوه (انظر: لسان العرب لابن منظور، ٩/٤٠١). وأما في العصر الحديث، فقد اندثر استخدام (عَمَلة) تقريبا، وتغيرت دلالة (عمال) فأصبحت تستخدم أكثر في الإشارة إلى الحرفيين من ذوي الدخل المنخفض وأصحاب الصناعات اليدوية والمهن الشاقة. وأما جمع التصحيح (عاملون) فيستخدم بمدلول عام للإشارة إلى جميع من يعمل في مجال معين من الموظفين وغيرهم. وهذا الاختلاف بين الكلمتين جار على الأصل لأن جمع التصحيح يدل غالبا على مطلق الوصف المناسب لدلالة مادة الاشتقاق، وأما جمع التكسير فيدل على تسمية مقصودة لصنف خاص(٣٨). وربما يوضح الكشاف السياقي لعينة



عشوائية من استخدام هاتين الكلمتين التفاوت في مدلوليهما كما في الجدولين (٣١) و(٣٢).

الأفضل فهذا ليس من شأن	العمال	المنفذين بل ربما يلتفت إليهم
الانتاج الالمانية الاسبوع الماضي إن	العمال	الصينيين الذي ينتجون الاحذية يحصلون
التجار ويسرق من تعب وعرق	العمال	وتابع نريد نقلا عاما أسوة
ان تستمر في تجاهل مطالب	العمال	ومطالب حركتهم النقابية وأعتبر المكتب
العمال وتمنى من الدولة إنصاف	العمال	والعائلات المحرومة بدلا من إنقاص
ومن لا يريد الدفاع عن	العمال	وحقهم فليرحل وقال الغلاء طال
وتحميل السفن وبالنسبة للكثير من	العمال	فإن عملهم صعب للغاية بسبب
ما يصعب كثيرا علينا وعلى	العمال	أثناء عمليات الحفر هذا إلى
ومن ثم لا يبقى أمام	العمال	سوى تركيب هذه الأجزاء وطلاء
النقابات حول ضرورة تأسيس اتحاد	العمال	الزراعيين في روسيا ص المصدر
بحيث يتم إعادة المسرحين من	العمال	إلى المصانع وإعادة تنشيط الحرف

الجدول (٣١) نموذج من الكشاف السياقي لكلمة (العمال) في المدونة

أو البعيدة يخرج لك كل	العاملين	في هذه المؤسسات موجهين كل
وهذا يدل علي ان ثلثي	العاملين	تقريبا لايحصلون علي تامين صحي
تذليلها وهذا هوالمطلوب وبما يشجع	العاملين	في هذا القطاع على عرض
النبي صلى الله عليه وسلم	العاملين	في الأعمال الخيرية والمساهمين سواء
ولبث روح الأمل في نفسية	العاملين	في حقل الدعوة يبشر سراقة
والمؤتمرات بصفة دائمة بهدف إطلاع	العاملين	بالوزارة على كل جديد وحديث
الأربعة أضعاف كما تضاعف عدد	العاملين	في هذا القطاع فازداد عددهم
في المائة من إجمالي عدد	العاملين	في دول مجلس التعاون في
خاصة أصحاب الدخول المنخفضة من	العاملين	في الدولة والقطاع الخاص والمعاشيين
أنها تستقطب النسبة العظمى من	العاملين	في القطاع الخاص في سورية
الدولية شاكرا في ختام كلمته	العاملين	في المؤسسة ومدارس التحفيظ مترحما

الجدول (٣٢) نموذج من الكشاف السياقي لكلمة (العاملين) في المدونة

ولو أردنا الاستمرار في مادة (ع م ل)، ولكن في الأفعال هذه المرة فسنجد أن الفعل (استعمل) يدل على عدد من المعاني. فهو على صيغة (استفعل)، والمعنى الأصلي لهذه الصيغة هو الطلب. فمن معاني (استعمل): سأله أن يعمل له أو طلب إليه العمل. ومن معاني هذه الصيغة أيضا الاتخاذ. ولذا فمن معاني (استعمل): جعله عاملا له. وقد تدل أيضا على اتخاذ الشيء أداة أو وسيلة يعمل بها. وهذا المعنى الأخير هو الباقي في الاستخدام المعاصر. وفي مقابل (استعمل) هناك فعل آخر على صيغة (استفعل) كان مقصورا في دلالته القديمة على معنى الطلب، ولم يكن يستخدم في معنى الاتخاذ، وهو الفعل (استخدم)، ولكن تحولت دلالته في الاستخدام المعاصر، فأصبح يستخدم في معنى الاتخاذ كمرادف لـ(استعمل)، كما يوضح أحد الباحثين:

«يقول المعربون في عصرنا مثلا (يستخدم هذا الدواء لاتقاء البرد)، والمراد هنا يستعمل. ويقال: (يستخدم هذا الفعل بمعنيين...)، بمعنى يستعمل. وكأن الفعل (استخدم) مرادف للفعل (استعمل). وإذا عدنا إلى فصيح العربية وجدنا الفعل (استخدم) بعيد عن معنى (استعمل). قال أهل العربية «استخدمه فأخدمه، أي استوهبه خادما فوهبه له. ويقال: اختدمت فلانا واستخدمته، أي سألته أن يخدمني». وعلى هذا كان الفعل لا يفارق سياق الخدمة. وهذا كله لا نجده في استعمال المعربين في عصرنا» [1]

ورغم أهمية هذا النص في التنبيه على الدلالة الجديدة لكلمة (استخدم) في العربية المعاصرة، ألا أنه لا يقدم أي معلومات عن مدى شيوع كل من الكلمتين في الاستعمال الفعلي، ولا مدى التطابق بين دلالتيهما في الاستخدام المعاصر. وبالبحث عن فعلي (استخدم) و(استعمل) ومصدريهما في مدونتنا فقد تبين أن استعمال (استخدم) أكثر بكثير من (استعمل) كما يوضح الجدول رقم (٣٣) (٣٩). وأما من ناحية الدلالة فهذا الفعلان يستعملان بمعنى واحد، يدل على

(١) السامرائي، ١٩٩٥، ص ٨٧

اتخاذ شيء ما أداة، ولا يوجد فرق بين دلالتهما كما توضح عينة عشوائية من الكشاف السياقي لكلا الفعلين في الجدولين (٣٤) و(٣٥).

الكلمة	استعمل	يستعمل	استعمال	استخدم	يستخدم	استخدام
العدد	٣٣	٤١	٢٤٣	٩٤	١٩٥	١٢٧٨
المجموع		٣١٧			١٥٦٧	

الجدول (٣٣) تكرار كلمتي (استعمل/استخدم) ومصدريهما في المدونة

العلماء الاختبارات الجينيه لتحديد الاف	استعمل	الى الجرعه العاديه منذ عام
مختلف اسلحه الدمار الشامل وساهموا	استعمل	نبي رغم انهم اول من
مضادات الحموضه التي تباع دون	استعمل	المعده كلما زادت كميه افرازاتها
وساده عاليه قليلا وذلك للتقليل	استعمل	والطريقه التي يصفها لك الصيدلي
عبارات اقل عنفا فان فكرته	استعمل	بحيث يتبين لو كان قد
عائلته كلها في سبيل الله	استعمل	بكر الصديق رضي الله عنه
الفرح بدل الغضب فقال الفرح	استعمل	الغضب الساطع ات ولكن الشاعر
المشاركون دراجه ثابته وساقوها بنشاط	استعمل	ففي الساعه الحاديه عشره صباحا
الاخر الغاز المسيل للدموع ورشه	استعمل	واحد ولم يكفهم ذلك بل
العون عصاه لضرب الشاب ورغم	استعمل	عليه وهنا بدا الخلاف حيث
الاخر الغاز المسيل للدموع ورشه	استعمل	واحد ولم يكفهم ذلك بل
الغاز المسيل للدموع لتفريقهم المباراه	استعمل	المشرفين على التنظيم والامن الذي

الجدول (٣٤) عينة من الكشاف السياقي لكلمة (استعمل)

طائرات من دون طيار بشكل	استخدم	الشرق الأوسط والمحيط الهادي أوباما
عبارات امرؤ القيس في معلقته	استخدم	واللافت في الأمر أن شاعرنا
الباحثون لائحة بيك وهو نظام	استخدم	طول أصابع الرجال والنساء ثم
هذا الأسلوب أفلا ينظرون إلى	استخدم	هذه واحدة من الطرق والقرآن
مناهج عدة منها على سبيل	استخدم	أفادته كثيرا في دراساته فهو
المفاتيح لأبوابها ومنهم عاند المفاتيح	استخدم	فترة إحراز المفاتيح فمن متخرج
الجيش قوته كحاكم للبلاد من	استخدم	مرسي يستخدم صلاحياته كرئيس مثلما
كل الأساليب الملتوية ففجأة يحصل	استخدم	لكي يبقى في السلطة لقد
الجنوبيون البنادق والحراب في قتل	استخدم	التمرد الدموي في اغسطس حيث
الفراعنه العسل فى التحنيط كما	استخدم	القدماء العسل كغذاء ودواء وقد
فيها نوعا من أنواع المونتاج	استخدم	في روايته الثانية المكلا التي

الجدول (٣٥) عينة من الكشاف السياقي لكلمة (استخدم)

مناقشة نتائج الدراسة

استعرضنا في هذا المبحث عدداً من الأمثلة التي توضح كيف يمكن أن تستخدم المدونات في البحث والتحليل اللغوي. وكما يظهر من النتائج السابقة فإن أهم ميزة للسانيات المدونات هو أنها تقدم منهجية اختبارية تعتمد على الأدوات الحاسوبية والتحليل الكمي الإحصائي لعينات وبيانات اللغة كما تستخدم فعلا في الواقع ومجالاته التواصلية المختلفة. ورغم أن هذا مهم جدا كما رأينا في استخراج نسب التكرار والشيوع، وتحول الدلالات، والمقارنة بين المترادفات، بالإضافة إلى التعرف إلى المفردات المميزة للمجالات التواصلية المختلفة، إلا أن هناك ملاحظات ينبغي الانتباه لها. منها أن لسانيات المدونات تركز بشكل أساسي على الجوانب المعجمية في اللغة. ولكن ذلك لا يعني استبعاد أو إهمال القضايا النحوية، فعلاقة المعجم بالتركيب علاقة متداخلة، ولها مستويات متعددة(٤٠). ومن نقاط الالتقاء المهمة بين المعجم والتركيب الكلمات الوظيفية التي تلعب دورا نحويا في المقام الأول عبر ربط مكونات الجملة لتكوين وأداء وظائف معينة.

وكما رأينا فإن الكلمات الوظيفية هي الأكثر بروزاً وتكراراً في المدونة. وهذه النتيجة ليست مفاجأة، بل هي المتوقعة في أي مدونة لغوية، فالكلمات الأكثر شيوعاً في كل اللغات تنتمي غالبا إلى الكلمات الوظيفية المغلقة (انظر مثلاً: Curzan & Adams, 2012, p. 102). ودون الدخول في تفاصيل دقيقة وكثيرة، فإن معاودة الوقوف على الكلمات الوظيفية بناء على نتائج لسانيات المدونات مهم جدا لسببين: أحدهما نظري، والآخر تطبيقي. أما النظري فهو أن يعاد النظر في موقع هذه الكلمات الوظيفية وأثرها في البناء اللغوي بناء على تصور نظري موحد وشامل. فقد اعتاد النحاة على تناول هذه المسألة بشكل موزع ومتفرق ضمن مباحث النحو العامة. ورغم أن بعضهم أفرد «حروف المعاني»

بالعناية والتأليف، إلا أن حروف المعاني أضيق مما نقصده هنا، لأن الكلمات الوظيفية تشتمل بالإضافة إلى حروف المعاني على «المبهمات» مثل الضمائر وأسماء الإشارة والأسماء الموصولة(٤١). كما تشتمل أيضا على «الأدوات المحولة» (حسان، ٢٠٠١، ص ١٢٣) التي تعد من قبيل الأدوات النحوية وتعمل عملها، ولكن النحاة لا يتناولونها ضمن حروف المعاني لأنها تنتمي إلى قسمي الأسماء أو الأفعال في التقسيم النحوي التقليدي لأقسام الكلام(٤٢).

وأما السبب التطبيقي فيتعلق بالمعالجة الإحصائية لهذه الأدوات كما تتحقق في واقع الاستخدام الفعلي. فغياب الدراسة الإحصائية جعل النحاة يتناولون في حروف المعاني حروفا نادرة الاستخدام أو منقرضة مثل (جير، جلل، عوض، بجل) على نفس المستوى مع حروف أخرى شائعة ومتكررة. كما جعلهم يتناولون أيضاً حروفاً مرتبطة غالباً بالمستوى الشفهي مثل (أجل)، أو بظواهر لهجية خاصة مثل (الميم، والشين) (انظر مثلا: الجنى الداني للمرادي ص ١٣٩و ١٥٢) مع غيرها بلا تمييز. ومن شأن الدراسة الإحصائية أن توضح نسبة تكرار الأدوات وشيوعها، ومواضع تكرارها واستخدامها. ومن ذلك مثلا أن حرف الجر (في) « كما رأينا » هو أكثر الحروف تكرارا في العربية بسبب تعدد معاني هذا الحرف وتداخله الدلالي مع عدد من الحروف الأخرى رغم أن معناه الأصلي هو الظرفية (انظر، مثلا: الشريف، ١٩٩٦، ٧٥١/٢). وقد تؤدي الدراسة الإحصائية أيضا إلى بيان بعض الاختلافات بين ما يقرره النحاة من أحكام وواقع الاستعمال الفعلي(٤٣) (انظر مثلا: الربطة، ٢٠٠٠). وهذا يتأكد في واقع الاستخدام اللغوي الحالي لـ«الفصحى المعاصرة». فكما رأينا، هناك أدوات وتعبيرات مثل (حول) و (من خلال) أصبحت ضمن الأكثر تكراراً بسبب التأثر بالترجمة والأساليب المعاصرة(٤٤).

ومن ناحية أخرى، فإن استخدام المدونات لا يقتصر على تتبع الأدوات والكلمات الوظيفية، بل هناك مجال واسع لاستخدامها في الدراسات المعجمية

المتعلقة بكلمات المحتوى. وهذا الجانب بالذات له ارتباط كبير بدراسة «الفصحى المعاصرة»، لأن من أبرز خصائصها تجاوز المعجم التراثي واستحداث كم هائل من الأبنية والدلالات والأساليب الجديدة. وكما رأينا فإن الصحافة مورد ثري للغاية يزخر بعدد وافر جدا من المفردات والدلالات الجديدة، فبعضها استحدث أساسا في الصحافة، وبعضها عملت الصحافة على إذاعته وانتشاره. ومع ذلك فإقصاء الصحافة لا يزال هو الموقف السائد في الدراسات اللغوية (الحمزاوي، ٢٠١١). ومن نتائج ذلك فقر المعاجم العربية الحديثة في مقابل المعاجم الاستشراقية التي اعتمدت إلى حد كبير على لغة الصحافة في تتبع المستحدث والمستجد في اللغة العربية (للاطلاع على بعض الأمثلة، انظر: كنون، ١٩٨٣؛ الحمزاوي، ٢٠٠٣). وحتى الذين يهتمون بهذا الجانب، يعتمدون على التتبع العشوائي لأمثلة جزئية ومتفرقة حسبما يتفق، دون الاستناد إلى منهجية إحصائية منضبطة. ومن نتائج ذلك اضطراب المعايير في اعتبار وإدراج المفردات والدلالات الجديدة في المعجم. ومن أمثلة ذلك - كما رأينا - خلو مادة «خ د م» في (المعجم الوسيط) من الدلالة الجديدة لكلمة «استخدم» كمرادف لـ«استعمل» رغم شيوعها وانتشارها الكبير، في مقابل إدراج كلمات جديدة أقل منها تكرارا بكثير مثل (زبون) و(عمولة).

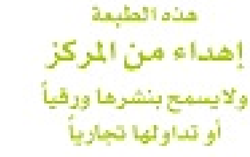
وعوداً على ما بدأنا به هذا المبحث، فإن مفهوم «الفصحى المعاصرة»، ومراحل تطور العربية بشكل عام لا يمكن تأسيسهما على الانطباعات الذاتية، والدراسات الجزئية المتفرقة. فالتناول الجاد لهذه المسائل يتطلب الاعتماد على الحصر الاستقصائي والإحصائي للظواهر اللغوية في كافة مجالات التداول اللغوي المختلفة، مع مقارنة تلك المجالات بعضها ببعض من ناحية، ومقارنتها بمراحل التطور التاريخي المختلفة للغة العربية من ناحية أخرى. ولسانيات المدونات تمثل آلية ومنهجية منضبطة ومناسبة جدا لتحقيق ذلك، ولكن ذلك يتطلب دراسات مؤسسية متتابعة ومتراكمة مع استخدام مدونات لغوية متنوعة

ومختلفة في محتوياتها وتمثيلها التاريخي. وربما يؤدي تراكم البحث في هذا الاتجاه إلى استخلاص تصورات واضحة ومستندة إلى نتائج اختبارية في مجال لا يزال يعاني من التعميمات الانطباعية بلا مستند واضح. والصحافة المكتوبة بالذات منذ مطلع النهضة العربية الحديثة تمثل وثيقة تاريخية مهمة جدا تحتوي على مسارات حركة اللغة العربية وتطورها خلال القرنين الماضيين. فدراستها بناء على المنهجية المستخدمة في لسانيات المدونة إضافة مهمة جداً، وانعدام ذلك لا يزال يمثل فجوة معرفية في الدراسات اللغوية العربية.

وآخر ما يمكن الإشارة إليه في هذا الصدد هو أن استخدام لسانيات المدونات في الدراسة اللغوية –رغم أهميته الكبيرة– له متطلباته الخاصة. وبرنامج (غواص) الذي استخدمناه في معالجة وتحليل بيانات الدراسة الحالية ليس إلا نموذجا للأدوات الحاسوبية التي يتطلبها العمل على لسانيات المدونات. ورغم أننا واجهنا بعض المشكلات في استخراج بعض الكلمات مثل الأدوات الملتصقة بالكلمات المصاحبة لها (مثل «السين» و«أل» التعريف)، والكلمات التي يتطلب التعرف على نوعها الخاص أن تكون مشكولة (مثل «مَن» الموصولة، و«مِن» حرف الجر، و«عَمِلَ» الفعل، و«عَمَل» المصدر) كما أشارنا إلى ذلك، إلا أنه كاف في تحقيق واستخراج النتائج التي كنا نصبو إليها في هذا الفصل. وهناك أنواع أخرى من التحليل اللغوي تتطلب مدونات موسومة وأدوات معالجة ذات وظائف إضافية. ولذا فإن لسانيات المدونات تتطلب التعاون الجاد والوثيق بين اللسانيين والحاسوبيين من أجل تطوير آليات التحليل والأدوات الحاسوبية المناسبة لمعالجة ظواهر وقضايا ومسائل اللغة العربية.

الخاتمة

لقد كان الهدف الأساسي لهذا المبحث هو عرض نماذج توضح بعض أساليب استخدام لسانيات المدونات في دراسة اللغة، وكان تركيزنا منصبا



على دراسة بعض الظواهر اللغوية في «الفصحى المعاصرة» عبر استعراض بعض النماذج والأمثلة لأنواع التحليل الممكنة في هذا الصدد. ورغم أننا قدمنا بعض النماذج مثل استخراج الكلمات الشائعة والكلمات المميزة والمقارنة بين بعض المفردات المترادفة، إلا أن هناك أنواعاً أخرى واستخدامات أخرى كثيرة للسانيات المدونات لم تتسع لها المساحة المتاحة. ومن الأمثلة المهمة ذات الارتباط بـ«العربية المعاصرة» التي كان يمكن أن نتناولها بالدراسة قضايا التعدية واللزوم واستخدام حروف الجر مع الأفعال، وقضايا الأساليب المستجدة، بالإضافة إلى الدراسات المتعلقة بتحليل الخطاب. ولكننا نكتفي بهذا القدر لأن المقصود هو تقديم نموذج فقط لنوعية المدونات، والأدوات الحاسوبية، وأنماط التحليل التي يمكن استخدامها في لسانيات المدونات.

الحواشي

(١٧) نود أن نشكر الزميل الدكتور منصور ميغري على تفضله بقراءة مسودة هذا المبحث، والملاحظات القيمة التي زودنا بها، كما نشكر الدكتور صالح العصيمي على متابعته ومراجعته لهذا الفصل وملاحظاته القيمة.

(١٨) حركة النهضة العربية ابتدأت في مطلع القرن التاسع عشر الميلادي، نتيجة لعوامل كثيرة ومتعددة، وتأثرت بها البلدان العربية بنسب متفاوتة ومختلفة، وقد كتب عنها الكثير من الناحية التاريخية والاجتماعية، وللمزيد حول بعض الآثار الفكرية واللغوية، انظر: حوراني، ١٩٦٨؛ الشيال، ١٩٥١.

(١٩) لابد من التنبيه هنا على أن هناك من يعارض مصطلح «العربية المعاصرة» (انظر على سبيل المثال: بشر، ١٩٩٩)، ويرى أن هذا المصطلح يوحي باختلاف العربية المعاصرة عن العربية في العصور

السابقة. ولكن كمال بشر نفسه الذي أطال في إنكار مصطلح «العربية المعاصرة» يقبل مصطلحا آخر مشابها له: «العربية في العصر الحديث»، ويدعو إليه، ولا ينكر اختلاف مستويات الاستخدام اللغوي، ولا وجود بعض السمات والملامح التعبيرية والأسلوبية التي تخص العربية في العصر الحديث (انظر: بشر، ١٩٩٩، ص ٥٩).

(٢٠) إبراهيم السامرائي، على سبيل المثال، من اللغويين العرب الذين كانت لهم عناية بهذا الموضوع، وله جهود بارزة في تتبع بعض الظواهر المستجدة في العربية، ويقول في دراسة مبكرة في هذا الصدد «وهذه أشتات جمعتها من هنا وهناك، ولم أتبع في جمعي هذا منهجا خاصا، فمنها ما شاع في لغة الصحافة اليومية، ومنها ما هو جار على ألسنة المذيعين، ومنها ما هو مستعمل في لغة الكتابة الخاصة، وأعني بالخاصة لغة الكتابة غير الأدبية كالألفاظ الاقتصادية والسياسية ونحو ذلك» (السامرائي، ١٩٧٣، ص ٥٧). ولا يزال هذا الأسلوب هو السائد والمتبع في دراسة ظواهر الاستعمال اللغوي في «الفصحى المعاصرة».

(٢١) السجلات/Registers، والضروب/Genres من مصطلحات اللسانيات واللسانيات الاجتماعية للتمييز بين مستويات وأنماط الاستخدام اللغوي المختلفة. ويختلف تعريفهما تبعا لاختلاف النظريات اللغوية. ولكن بشكل عام، تستخدم «السجلات» للإشارة إلى ارتباطات النصوص بالمجالات والسياقات الاجتماعية المختلفة، وأما «الضروب» فتستخدم للإشارة إلى فئات النصوص المختلفة وما تشتمل عليه من انتظامات داخلية وشكلية تمنحها هويتها المميزة (وللمزيد، انظر: Biber & Conrad, 2009؛Lee, 2001).

(٢٢) لابد من التنبيه هنا إلى أن المكانة النظرية للسانيات المدونات مسألة خلافية. ففي حين ينظر لها الأغلبية على أنها مجرد منهجية بحث

اختبارية، يرى بعض الباحثين (Tognini-Bonelli, 2001) أنها لا تخلو من منطلقات نظرية تخصها، بل يذهب بعض الباحثين (Sinclair, 2004) أبعد من ذلك، ويعتبر أن لسانيات المدونات بذاتها تمثل إطارا نظريا جديدا وبديلا للبحث اللغوي. وقد ناقش ماكنري وغابريلاتوس (McEnery & Gabrielatos, 2006) هذه المسألة بتفصيل أوسع.

(٢٣) نفترض هنا أن القارئ، بناء على ما قُدِم في المباحث السابقة في هذا الكتاب، على وعي ودراية بأنواع التحليل الممكنة والمستخدمة في لسانيات المدونات، ولذا فإننا لن ندخل في بيان التفاصيل في هذا الصدد (وللمزيد يمكن الرجوع إلى: صالح، ٢٠١٢؛ العصيمي، ٢٠١٣؛ وأيضا: Hunston, 2006).

(٢٤) ونود أن نوضح هنا أنه يوجد اختلاف طفيف في حساب عدد كلمات المدونة بين ما ذكرناه هنا، وما هو وارد في الورقة العلمية التي نشرت عن المدونة سابقا (Al-thubaity et al., 2013)، حيث تشير الورقة إلى أن عدد كلمات المدونة هو ٢،٢٠٧،٤٦٩، في حين أن ما وجدناه هو ٢،١٨٤،٣٧٥، أي بفارق ٢٣،٠٩٤. ويعود السبب في ذلك إلى استخدامنا لنسخة محدثة من غواص حسّنت من طريقة حساب كلمات المدونة.

(٢٥) أدت الاضطرابات السياسية في بلدان مثل ليبيا وسوريا إلى تعطل الكثير من مواقع الصحف مما أثر بشكل واضح على عدد النصوص المجموعة من تلك الدول.

(٢٦) رغم أنه يمكن استعراض جميع الكلمات الواردة في المدونة إلا أننا سنكتفي هنا بالحديث بشكل موجز عن الكلمات المئة الأكثر تكرارا سواء في المدونة كاملة أو في أجزائها المقسمة على المجالات المختلفة.

(٢٧) هذه المصطلحات متداخلة وتستخدم بمدلولات مختلفة، ولكن «الكلمات الوظيفية»، وهي كلمات تنتمي إلى فئة مغلقة، تؤدي وظائف نحوية في المقام الأول، وليس لها دلالة مرجعية غالبا، أوسع من «الأدوات»، لأنها تشتمل على «الأدوات» وما يشبهها كالضمائر وأسماء الإشارة والموصولات. و«الأدوات» أشمل مما اعتيد على تسميته بـ«حروف المعاني» في التراث النحوي (انظر مثلا: الساقي، ١٩٧٧، ص ٢٦٢). وسنتوقف لتوضيح ذلك بشكل موجز في الجزء المخصص لمناقشة نتائج الدراسة.

(٢٨) ولذا فإن نحاة أهل الكوفة يجعلون الأسماء المبهمة التي تلازم الظرفية والإضافة غالبا من قبيل حروف الخفض كما يشير ابن السراج مثلا (انظر: الأصول لابن السراج ١/٢٠٤). ورغم أنه يذكر ذلك في معرض النقد والتشنيع عليهم بسبب عدم تفريقهم بين الحروف والأسماء، إلا أن ذلك يدل على إحساسهم بمدى قرب هذه الكلمات من الحروف والكلمات والوظيفية. وشبيه بذلك ما نسب إلى سيبويه ونحاة آخرين أنهم يجعلون (على) اسم ظرف، ولا يعدونها ضمن حروف الجر (انظر: البحر المحيط لأبي حيان، ١/١٤٥). وهذا جزء من اختلافات طويلة ومتشعبة في النحو العربي حول انتماء عدد كبير من الكلمات إلى «أقسام الكلم» المختلفة. ورغم أننا لن نتطرق إلى هذا الموضوع، إلا أننا سنشير إلى ما يترتب على ذلك من مقتضيات في الجزء المخصص لمناقشة نتائج الدراسة.

(٢٩) لابد من التنبيه إلى أنه بسبب عدم عرض الحركات فإن «من» هنا قد تكون «مِن» مكسورة الميم التي هي حرف جر، أو «مَن» مفتوحة الميم التي تكون اسما موصولا أو اسم استفهام. ويمكن التحقق من ذلك عبر الكشاف السياقي، ولكن ذلك خارج عن مقصودنا هنا.

(٣٠) لابد من التنبيه هنا إلى أن حروف المعاني الأحادية والملتصقة بالكلمات اللاحقة لا تظهر في التحليل بسبب اعتبارها جزءا من الكلمة، ولذا فإن واو العطف، و(أل) التعريف اللذين يتوقع أن يكونا من أكثر الأدوات تكرارا لم يظهرا كما هو متوقع. ورغم وجود برامج تتيح التعامل مع هذه المشكلة، إلا أننا لم نستخدمها لسببين، أولا عدم دقتها في بعض المواضع، والثاني خروج ذلك عن مقصودنا الأساسي، وهو الاكتفاء بعرض بعض النماذج والعينات.

(٣١) لاحظ أن «الظروف» كلها تتضمن معنى «في» باطراد (انظر مثلا: حسن،١٩٧٣، ٢٤٢/٢؛ السامرائي، ٢٠٠٠، ١٧٨/٢).

(٣٢) ومن ناحية أخرى، يرتبط هذا الفعل بمبحث أساسي في التركيب المعاصر، وهو أفعال العماد/Support verbs أو «الأفعال الناقلة» وهي أفعال من قبيل الحشو لأنها فارغة دلاليا تقريبا، ولا تمثل ركنا في الإسناد، بل يقتصر دورها الأساسي على الوصل بين ركني الإسناد في الجملة. ويعد شيوع استعمالها من الخصائص المميزة في العربية المعاصرة (انظر للمزيد: Ashtiany، ١٩٩٣؛ الورهاني، ٢٠٠٨).

(٣٣) ومن المعربات القديمة التي وردت ضمن الكلمات المميزة في هذا المجال كلمة (رصيد)، وهي من الكلمات التي توضح خفاء حركة الاقتراض اللغوي في بعض الأحيان (انظر: السغروشني، ١٩٩٦).

(٣٤) لابد من التنبيه هنا إلى أن الحكم بأن هذه الأبنية جديدة ليس أمرا يستنبط من المدونة، إلا إذا كانت مدونة تاريخية تتيح تتبع تطور الألفاظ عبر المراحل الزمنية المختلفة، ومدونتنا ليست كذلك. ولذا فإن الحكم بأن هذه الأبنية أبنية مستحدثة حكم مسبق اعتمدنا فيه على المعاجم الحديثة مثل (المعجم الوسيط)، وعلى كلام بعض الباحثين مثل

الدكتور إبراهيم السامرائي (١٩٧٣) الذي يقول في كلمة «عميل» مثلا «لا يوجد في العربية صفة «فعيل» من هذه المادة، فالكلمة في صورتها الاشتقاقية جديدة، ومعناها جديد أيضا» (ص ٦٣).

(٣٥) هذه الرموز تشير إلى المجالات الفرعية في المدونة فـ(ثقف) تشير إلى المجال الثقافي...وهكذا، و(كمل) تشير إلى المدونة كاملة.

(٣٦) المصدر على وزن (فعولة) الأكثر على أنه سماعي، وبعضهم يرى أنه قياسي في باب (فَعُل، يَفعُل) مثل (صَعُب، يَصعُب، صعوبة). ولكن يلاحظ أنه مر معنا مصدران جديدان مبينان على هذه الصيغة وليسا من هذا الباب (سيولة، عمولة).

(٣٧) ورد في لسان العرب لابن منظور (١٩٩٩، ٣/١٣٠) «وفلان حريفي أي معاملي....وحريف الرجل: معامله في حرفته» (...)، وانظر أيضا (السامرائي، ١٩٧٣، ص ٧٢).

(٣٨) هذه نقطة في غاية الأهمية، ولا يتسع المجال للتفصيل، وللمزيد حول هذا الفرق الدلالي بين جموع الصحة وجموع التكسير، انظر: السامرائي، ٢٠٠٧، ص ١٢٧. وللدكتور محمود الطناحي (١٩٩٢) دراسة قيمة في بيان أثر العرف اللغوي في تقييد دلالات أبنية جموع التكسير، ومنه استوحينا فكرة المقارنة بين جموع (عامل). واختلاف دلالة أوزان الجموع من الأبواب التي تحتاج إلى دراسة ولم تحظ بعناية كبيرة (انظر مثلا: السامرائي، ٢٠٠٧، ص ١١٣). فبعض الأوزان تختص بمجال دلالي معين مثل كلمة (عباد) التي تختص « كما مر منا في الجدول (٢٠) » بالمجال الديني، بخلاف كلمة (عبيد)، وكلاهما جمع تكسير لكلمة (عبد).

(٣٩) ومع ذلك لم يشر المعجم الوسيط إلى هذه الدلالة الحديثة لكلمة (استخدم) رغم عنايته بما طرأ في العربية من كلمات محدثة ومولدة، ولكن هذا المعنى موجود في (معجم اللغة العربية المعاصرة) لأحمد مختار عمر (٢٠٠٨).

(٤٠) لابد من التنبيه هنا إلى أن النظريات اللسانية تختلف فيما بينها في تصور طبيعة العلاقة بين النحو والمعجم. ففي حين أن النظرية التوليدية، خصوصا في مراحلها الأولى، تؤكد على أسبقية واستقلالية التركيب بمعزل عن المعنى، نجد أن النظرية المنظومية، لمايكل هالدَي ومن يتابعه، تشدد على وحدة النحو والمعجم واعتبارهما دعامتين لبناء واحد «the lexico-grammar cline» (انظر: Halliday & Mattiesssen، ٢٠٠٤، -. ٤٣).

(٤١) وكما رأينا سابقا فإن الظروف المبهمة التي تلازم الإضافة غالبا أو دائما مثل «بين، قبل، بعد، دون...» من قبيل المبهمات، وتندرج ضن الكلمات ذات التكرار العالي. ورغم كونها من قبيل كلمات المحتوى، إلا أن إبهامها الدلالي المتأصل يجعلها قريبة جدا من الكلمات الوظيفية (وانظر أيضا: كامل، ٢٠١٣).

(٤٢) كما هو واضح فإن هذه المسألة مرتبطة بتصنيف أقسام الكلام في النحو العربي. والتقسيم الثلاثي التقليدي لأقسام الكلام: اسم، وفعل، وحرف، مجرد جدا، ولا يفي بمعالجة خصوصيات الكلمات الوظيفية (غيوم، ١٩٩٦)، ولذا وقع كثير من الاضطراب في تصنيف كثير من الكلمات الوظيفية ضمن أحد هذه الأقسام (لمناقشة مستفيضة لهذه المسألة، انظر: الساقي، ١٩٧٧). ورغم وجود عدد من المحاولات الحديثة لوضع نماذج بديلة لتصنيف أقسام الكلام (للاطلاع على بعضها انظر: عاشور، ٢٠٠٥؛ كامل، ٢٠١٣)، ومن أبرزها محاولة تمام

حسان (حسان، ٢٠٠١) وتلميذه فاضل الساقي (١٩٧٧)، إلا أنه لم يكتب لها الشيوع والذيوع الكافي.

(٤٣) ومن الأمثلة المشهورة في الاستدراك على بعض التقريرات النحوية بناء على الاستقراء الإحصائي ما فعله عبدالخالق عضيمة في كتابه «دراسات لأسلوب القرآن» (انظر مثلا: عيسى وعلي، ٢٠٠٩).

(٤٤) ومن الأمثلة البارزة على أدوات استجد لها استخدامات حديثة في «العربية المعاصرة» كلمة «أيّ» حيث أصبحت تستخدم كأداة ربط مقابلة لكلمة 'any' الإنجليزية، ولم تكن تستخدم بهذه الطريقة سابقا. ولكننا لم نتناول هذه المسألة في الدراسة الحالية لضيق المجال.

المراجع

المراجع العربية:

أبو الهيجاء، ياسين (٢٠١٠) إشكالية تعريب الأساليب في قرارات لجنة الألفاظ والأساليب في مجمع اللغة العربية بالقاهرة: الرابط نموذجا، المجلة الأردنية في اللغة العربية وآدابها، م (٦)، ع (١)، ٩٥-١٢٠.

أبوحيان الأندلسي، محمد بن يوسف (١٩٩٣) تفسير البحر المحيط، تحقيق: عادل أحمد عبدالموجود وعلي محمد معوض، بيروت: دار الكتب العلمية.

بدوي، السعيد محمد (٢٠١٢) مستويات العربية المعاصرة في مصر: بحث في علاقة اللغة بالحضارة. القاهرة: دار السلام للطباعة والنشر والتوزيع.

بشر، كمال (١٩٩٩) اللغة العربية بين الوهم وسوء الفهم. القاهرة: دار غريب للطباعة والنشر والتوزيع.

البعلبكي، منير (١٩٨٨) الإعلام واللغة الإعلامية، مجلة مجمع اللغة العربية، القاهرة، عدد (٦٢)، ص ٢١٠-٢٤٤.

البلداوي، حسن جعفر (٢٠١٢) التعدية بالحروف في المعجم الوجيز، مجلة الأستاذ (كلية التربية، جامعة بغداد)، ع (٢٠٢)، ص ٢٥٥-٢٦٤.

بيلكين، ف. م (١٩٧٣) في تاريخ تطور اللغة العربية الفصحى، ترجمة: جليل كمال الدين، مجلة المورد (وزارة الإعلام العراقية)، م (٢)، ع (١)، ص ٣٣-٣٩.

حسان، تمام (٢٠٠١) اللغة العربية: معناها ومبناها، الدار البيضاء، دار الثقافة.

حسن، عباس (١٩٧٣) النحو الوافي، القاهرة: دار المعارف.

حمادي، محمد ضاري (١٩٨١) حركة التصحيح اللغوي في العصر الحديث. بغداد: مطبوعات وزارة الإعلام والثقافة العراقية، دار الرشيد.

حمادي، محمد ضاري (١٩٩٩) التعدية بالحرف «على» في تحقيقات اللغويين، مجلة المجمع العلمي العراقي، م (٤٦)، ج (٢)، ص ٨٧-١٠٢.

الحمزاوي، محمد رشاد (١٩٨٦) العربية والحداثة: أو الفصاحة فصاحات. بيروت: دار الغرب الإسلامي.

الحمزاوي، محمد رشاد (٢٠٠٣) في لغة الصحافة وتعريب العلوم: قضاياها، وإشكالاتها، ومناهج دراستها، مجلة اللسان العربي، مكتب تنسيق التعريب، عدد (٥٦)، يمكن الرجوع إليها من خلال الرابط التالي: http://www.arabization.org.ma/hsearch.aspx

حوراني، البرت (١٩٦٨) الفكر العربي في عصر النهضة ١٧٩٨-١٩٣٩، ترجمة: كريم عزقول. بيروت: دار النهار للنشر. (١٩٦١).

خلف، ربيع عبدالسلام (٢٠١١) التركيب الموسع في الفصحى المعاصرة. المجلة العربية للعلوم الإنسانية (جامعة الكويت)، م (٢٩)، ع (١١٦).

الربطة، عادل (٢٠٠٠) حروف المعاني في الاستعمال الجاري: مثل من القديم والحديث، رسالة ماجستير غير مطبوعة، الجامعة الأردنية.

الزعبي، آمنة صالح (٢٠٠٦) في تحول الأساليب النحوية (التركيبية) في اللغة العربية في العقدين السابقين على مرحلة العولمة: لغة القصة القصيرة في الأردن نموذجا، مجلة جامعة دمشق، م (٢٢)، ج (١-٢)، ١٣١-١٦٤.

الساقي، فاضل مصطفى (١٩٧٧) أقسام الكلام العربي من حيث الشكل والوظيفة، القاهرة، مكتبة الخانجي.

السامرائي، إبراهيم (١٩٧٣) تنمية اللغة العربية في العصر الحديث، المنظمة العربية للتربية والثقافة والعلوم: معهد البحوث والدراسات العربية.

السامرائي، إبراهيم (١٩٩٥) في العربية المعاصرة ومعجمها، مجلة مجمع اللغة العربية، القاهرة، عدد (٧٦)، ص ٧٨-٩٦.

السامرائي، إبراهيم (٢٠٠٠) تصحيح «التصحيح»، مجلة مجمع اللغة العربية، القاهرة، عدد (٨٧)، القسم الأول، ١٣٣-١٤٣.

السامرائي، فاضل (٢٠٠٠) معاني النحو، عمان: دار الفكر للطباعة والنشر والتوزيع.

السامرائي، فاضل (٢٠٠٧) معاني الأبنية في العربية، عمان: دار عمار للنشر والتوزيع.

ستتكيفتش، جاروسلاف (١٩٨٥) العربية الفصحى الحديثة: بحوث في تطور الألفاظ والأساليب، القاهرة: دار النمر للطباعة.

(ابن) السراج، محمد بن سهل (١٩٩٦) الأصول في النحو (ط. ٣)، تحقيق: عبدالحسين الفتلي، بيروت: مؤسسة الرسالة.

السغروشني، إدريس (١٩٩٦) حول الاقتراض، ضمن: عبدالقادر الفاسي الفهري (مح.)، اللسانيات المقارنة واللغات في المغرب، الرباط: كلية الآداب والعلوم الإنسانية.

سمبس، أميرة زبير (٢٠٠٦/١٤٢٧هـ) السمات اللغوية في صحيفة أم القرى في ضوء إسهامها الإعلامي والإداري في الفترتين (١٤٤٣-١٤٧٣هـ/١٤٠٢-١٤٢٣هـ). مجلة جامعة أم القرى لعلوم الشريعة واللغة العربية وآدابها، ج (١٨)، ع (٣٩)، ص ٣٢٣-٣٧٤.

السوسوة، عباس (٢٠٠٢) العربية الفصحى المعاصرة وأصولها التراثية، القاهرة: دار غريب.

الشريف، محمد حسن (١٩٩٦) معجم حروف المعاني في القرآن الكريم: مفهوم شامل مع تحديد دلالة الأدوات، بيروت: مؤسسة الرسالة.

الشيال، جمال الدين (١٩٥١) تاريخ الترجمة والحركة الثقافية في عصر محمد علي. القاهرة: دار الفكر.

صالح، محمود إسماعيل (٢٠١٢) الحاسوب والبحث اللغوي: المدونات اللغوية نموذجا، الرياض: جامعة الأميرة نورة بنت عبدالرحمن (منشورات كرسي بحث صحيفة الجزيرة للدراسات اللغوية الحديثة).

الطناحي، محمود (١٩٩٢) جموع التكسير والعرف اللغوي، مجلة مجمع اللغة العربية، القاهرة، عدد (٧١)، ص ١٣٩-٢١٢.

عاشور، المنصف (٢٠٠٥) دروس في أصول النظرية النحوية العربية من السمات إلى المقولات أو لولبية الوسم الموضعي، تونس: مركز النشر الجامعي.

عبدالعزيز، محمد حسن (١٩٨٧) لغة الصحافة المعاصرة، القاهرة: دار المعارف.

عبدالعزيز، محمد حسن (١٩٩٨) خصائص العربية المعاصرة: مظاهرة حداثتها في المفردات والتراكيب، مجلة اللسان العربي، مكتب تنسيق التعريب، عدد (٤٥)، يمكن الرجوع إليها من خلال الرابط التالي: http://www.arabization.org.ma/hsearch.aspx

عبدالكريم، جمعان (٢٠٠٨) اللغة العربية الفصيحة المعاصرة: محاولة لمقاربة المصطلح والمفهوم، مجلة اللسان العربي، مكتب تنسيق التعريب، عدد (٦١)، يمكن الرجوع إليها من خلال الرابط التالي: http://www.arabization.org.ma/hsearch.aspx

عصفور، محمد حسن (٢٠٠٤) تأثير الترجمة على اللغة العربية، مجلة جامعة الشارقة للعلوم الشرعية والإنسانية، م (٤)، ع (٢)، ١٩٥-٢١٦.

العصيمي، صالح (٢٠١٣) لسانيات المتون وعلوم اللغة، مجلة كلية الآداب والعلوم الإنسانية (جامعة سيدي محمد بن عبدالله، المغرب)، ع (١٩)، ص ٣٧-٦٧.

عمر، أحمد مختار (١٩٩٣) أخطاء اللغة العربية المعاصرة عند الكتاب والإذاعيين (ط.٢)، القاهرة: دار عالم الكتب.

عمر، أحمد مختار (٢٠٠٨) معجم اللغة العربية المعاصرة، القاهرة، عالم الكتب.

عيسى، خير الدين، وعلي، عماد (٢٠٠٩) ردود عبدالخالق عضيمة على النحاة في كتابه «دراسات لأسلوب القرآن الكريم»، مجلة جامعة كركوك للدراسات الإنسانية، عدد (١)، مج(٤)، ص ٩٧-١٢٨.

غيوم، جان باتريك (١٩٩٦) تكوين نظرية أقسام الكلام وبناؤها في العرف النحوي العربي، ترجمة: سام عمار، مجلة التعريب، عدد (١٢)، ص ٦١-٧٩. (١٩٨٨).

فايد، وفاء كامل (٢٠٠٣) بعض صور التعبيرات الاصطلاحية في العربية المعاصرة، مجلة مجمع اللغة العربية بدمشق، م (٧٨)، ج (٤)، ص ٨٩٥-٩١٦.

فرستيغ، كيس (٢٠٠٣) اللغة العربية: تاريخها، ومستوياتها، وتأثيراتها، ترجمة: محمد الشرقاوي، القاهرة: المجلس الأعلى للثقافة، المشروع القومي للترجمة. (١٩٩١).

فريمان، أندرو (٢٠١٣) طبيعة اللغة العربية القديمة وتغيرها إلى العربية الوسيطة ومن ثَمّ إلى العربية المعاصرة، ترجمة: حمزة المزيني، ضمن كتاب «دراسات في تاريخ اللغة العربية»، الأردن، عمان: دار كنوز المعرفة العلمية. (١٩٩٥).

فضل، عاطف (٢٠١٠) تمثلات المنهج الوصفي الإحصائي في الدراسات اللغوي الحديثة، مجلة التربية والتعليم (جامعة الموصل)، م (١٧)، ع (٤)، ص ١٨٥-٢٠٧.

القاعود، حلمي محمد (٢٠٠٨) تطور النثر العربي في العصر الحديث، الرياض: دار النشر الدولي.

القضماني، رضوان؛ عبدالقادر، ميساء (٢٠٠٤) التوليد اللغوي في المعاجم عند المحدثين، مجلة جامعة تشرين للدراسات والبحوث العلمية (اللاذقية)، م (٢٦)، ع (١).

كامل، سلاف مصطفى (٢٠١٣) أقسام الكلم في ضوء النظرية المعجمية الحديثة، مجلة الأستاذ (كلية التربية، جامعة بغداد)، عدد (٢٠٦)، ص ١٥-٤٠.

كنون، عبدالله (١٩٨٣) الصحافة وتجديد اللغة، مجلة مجمع اللغة العربية القاهرة، ج (٥١)، ص ١٢٥-١٣٢.

مجمع اللغة العربية (١٩٨٣) في أصول اللغة ج٣ (ط٣)، القاهرة: الهيئة العامة لشؤون المطابع الأميرية.

مجمع اللغة العربية (١٩٨٩) القرارات المجمعية في الألفاظ والأساليب من ١٩٣٤ إلى ١٩٨٧، القاهرة: الهيئة العامة لشؤون المطابع الأميرية.

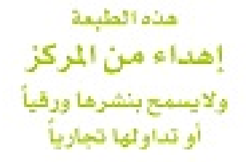
مجمع اللغة العربية (٢٠٠٤) المعجم الوسيط (ط. ٤)، القاهرة: مكتبة الشروق الدولية.

المرادي، الحسن بن قاسم (١٩٩٢) الجنى الداني في حروف المعاني، تحقيق: فخر الدين قباوة ومحمد نديم فاضل، بيروت: دار الكتب العلمية.

(ابن) منظور، محمد بن مكرم (١٩٩٩) لسان العرب، تحقيق: أمين محمد عبدالوهاب ومحمد الصادق العبيدي، بيروت: دار إحياء التراث العربي.

نصار، جهاد عبدالقادر؛ حماد، خليل عبدالفتاح (٢٠١٤) استعمالات لغوية معاصرة: دراسة في تأصيل الوضع اللغوي والتسويغ، مجلة جامعة فلسطين للأبحاث والدراسات، ع (٦)، ص ٢٨٣-٣٠٢.

الورهاني، بشير (٢٠٠٨) الأفعال الناقلة في العربية المعاصرة: بحث في الخصائص التركيبية والدلالية، تونس: المطبعة الرسمية للجمهورية التونسية.



المراجع الإنجليزية:

Al-Thubaity, A.; et al. (2013). New Language Resources for Arabic: Corpus Containing More Than Two Million Words and a Corpus Processing Tool. «Asian Language Processing (IALP), International Conference on. IEEE.

Ashtiany, J. (1993). Media Arabic. Edinburgh: Edinburgh University Press.

Bassiouney, R. (2009). Arabic sociolinguistics. Edinburgh: Edinburgh University Press.

Bateson, M. (2003). Arabic language handbook. Washington, DC: Georgetown University Press.

Biber, D. & Conrad, S. (2009). Register, Genre and style. Cambridge: Cambridge University Press.

Curzan, A. & Adams, M. (2012). How English works: A linguistic introduction. Canada: Person.

Halliday, M. & Matthiessen, C. (2004). An introduction to functional grammar (3rd ed.). London: Arnold.

Hunston, S. (2006). Corpus linguistics. Linguistics, 7(2), 215-244.

Lee, D. Y. (2001). Genres, registers, text types, domains and styles: Clarify the concepts and navigating a path through the BNC jungle. Language Learning & Technology, 5(3), 37-72.

McEnery, T., & Gabrielatos, C. (2006). English Corpus Linguistics. In B. Aarts & A. McMahon (ed.), The handbook of English linguistics. MA, Malden: Blackwell Publishing.

Ryding, K. (2005). A reference grammar of Modern Standard Arabic. Cambridge: Cambridge University Press.

Sinclair, J. (2004). Trust the text: Language, corpus and discourse. London: Routledge.

Tognini-Bonelli, E. (2001). Corpus linguistics at work. Amsterdam: Benjamins.

المبحث الخامس

البحث اللغوي في المدونات العربية الحاسوبية
بين الممكن والمحتمل والمأمول

سلطان بن ناصر المجيول

قسم اللغة العربية - جامعة الملك سعود

تمهيد

ليس ثمة مناص من التداخل الاختصاصي interdisciplinary بين الحقول اللسانية والعلوم الأخرى في الدوائر المعرفية والعلمية بحثًا وتحليلًا للسانيين والحاسوبيين والمحللين للخطاب اللغوي. ومن أوجه هذا التداخل الاختصاصي: ولوج تقنيات الحاسب وبرمجياته وأدواته التحليلية والمعالجاتية في جمع النصوص في مدونات محوسبة، وتصنيفها وفق أوعيتها المعلوماتية الناقلة، وفحصها تكرارًا وجمعًا وفرزًا وتحليلًا من مكانزها الرقمية. ومن المهم البدء بأهمية البحث اللغوي الحاسوبي وقضاياه الأولية، والتي لا يمكن حصر كل مناهجه، غير أن ثمة مفاتيح أساسية للباحث اللغوي تُمكنه من البدء في الدخول إلى هذا المجال، ويوظف عن طريقها خبرته اللغوية في أثناء معالجة قضية لغوية في المدونات الحاسوبية الأحادية (انظر على سبيل المثال، ميكنري وهاردي McEnery and Hardie ٢٠١٢؛ وبيكر Baker ٢٠١٠) أو التقابلية (جرانجر Granger ٢٠٠٨؛ وجوهانسون Johansson ٢٠٠٨) أو الترجمية (جرانجر وآخرون Granger et al ٢٠٠٨). ومن الممكن أن يقال إن ما سيقدم هنا هو ما يستحثه واقع البحث اللغوي العربي الحاسوبي، وبمحاولة جديدة في الدراسات اللغويات الحاسوبية للَملمة مناهج البحث اللغوي الحاسوبي وعرضها بصورة موجزة لا تقتصر على التعريف والعرض فحسب، بل على فتح أبواب البحث اللغوي الحاسوبي التطبيقي بصورة لعلها توضح معالمها للمختصين في اللغويات والدراسات اللغوية. وتُعرّف هذه الورقة بطرائق البحث اللغوي العربي الحاسوبي بصورة أولية تَرسم تمرحلات التحليل بواسطة المعالجة الحاسوبية بشكل مبسط؛ بدءًا بالتعريف بأنواع المدونات الحاسوبية للغة الطبيعية natural language، واتجاهاتها الممكنة والمحتملة والمأمولة، كل ذلك من أجل فحص المعجم العربي الذهني التوليدي الواقعي الذي يتمثل في المدونات الحاسوبية والبحث اللغوي من حيث

اتساع هذا المعجم أو تحجيمه باتساع المدونة أو تحجيمها (علاقة البحث اللغوي بتمثيل مدونة ما the representativeness of a particular corpus للمستويات اللغوية العربية النموذجية)، مرورًا بخصائص كل طريقة ومناهجها البحثية من حيث الغرض والتحليل والمعالجة، وصولًا إلى اقتراح بعضٍ من الموضوعات البحثية المهمة في هذا المجال.

البحث اللغوي والمدونات العربية الحاسوبية

تنال الدراسات اللغوية الحاسوبية الغربية -وإرهاصات هذا الدرس في سياق اللغة العربية والحاسب والمعالجة الآلية والبحث اللغوي القائم عليها- حظاً من التداخلات الاختصاصية بالقدر الذي يحتاجه التطوير في طرائق البحث اللغوي العربي الحاسوبي، وهذه التقاطعات هي على النحو التالي: أولًا: هندسة اللغة language engineering: (استعمال أدوات التحليل الحاسوبي لمعالجة اللغة وهندستها من حيث عناصر بنائها وعلاقة هذه العناصر في علم الوجود ontology). ثانيًا: اللسانيات الحاسوبية computational linguistics ومجاله الاختصاصي المعروف بمعالجة اللغة الطبيعية natural language processing NLP: (طرائق تأسيس وبناء خوارزميات وبرامج حاسوبية تُميكن اللغة وتسعى بأدبيات متعددة ومعقدة ومتنوعة إلى محاولات تعريف الخصائص اللغوية النوعية مع الخصائص الحاسوبية الكمية والرقمية). ثالثًا: لسانيات المدونات corpus linguistics: بناء مدونات حاسوبية، وإجراء بحوث لسانية حاسوبية إحصائية تفيد مجال تقويم مصادر اللغة Language Resources LRs، بالإضافة إلى التوسيم الصرفي والنحوي والدلالي والإحالي وتحشية النصوص المكتوبة أو المنطوقة في قوائم حاسوبية تحليلية، وإعدادها لمرحلتي التدريب training والاختبار test: أي مرحلة الإعداد والتحشية والتوسيم (التدريب)، ومرحلة اختبار عمليات التوسيم على القوائم من أجل تحقيق نسبة مئوية تقل

بازدياد عدد الكلمات في أي مدونة وتزيد بقلة عدد الكلمات في أي مدونة. رابعًا: تحليل النصوص الإلكترونية electronic text analysis: (جمع النصوص، والتنقيح، والمعالجة، والإخراج في قوائم حاسوبية من أجل التحليل لغرض معين، أو تضمينها في مواقع الويب، أو إدخالها في أقراص ممغنطة، إلخ). وسيُعرّف هذا المبحث الاختصاصيين بالدرس اللساني الحديث على حقل لسانيات المدونات الحاسوبية corpus linguistics أو علم المتون دون غيرها (انظر العصيمي ٢٠١٣ في مناقشة المقابلات العربية لمصطلح corpus linguistics).

وثمة أسئلة عديدة سيُجاب عنها من خلال هذه الورقة في الجملة، ومن أهمها: هل من الممكن أن تُجيبنا المدونات عن كل أسئلة البحث اللغوي؟ وهل للبحث اللغوي المعتمِد على المدونات شروط منهجية؟ وأي مدونة؟ ولماذا؟ إن من أبرز المسارات البحثية التطبيقية للدرس اللساني الحاسوبي والتحليلي في المدونات الحاسوبية اللغوية التنوّع الوصفي synchronic variation، والتنوّع الزمني diachronic variation، ومباحث التغير اللغوي الوصفي والزمني ضمن المجالات اللسانية التركيبية والدلالية، والتغير اللغوي الوصفي في العربية المعاصرة، وذاك الزمني في التاريخ التحليلي historiography للغة العربية.

وسيُتطرق في هذا المبحث إلى عدة موضوعات؛ أولها: الفرق بين الدرس اللساني المعتمد على المدونة corpus-based وذلك الموجه بالمدونة corpus-driven. وثانيها: أنواع المدونات الحاسوبية من حيث عدد اللغة فيها إلى أحادية اللغة monolingual وتعددية اللغة multilingual، وتنقسم الأخيرة إلى تقابلية comparable (انظر: شاروف Sharoff ٢٠٠٤؛ والمجيول al-Mujaiwel ٢٠١٢)، ومتوازية parallel[٤٥] (انظر العجمي al-Ajmi ٢٠٠٣). وثالثها: الفرق بين الأدوات التحليلية للغة العربية ودورها في قراءة الرموز العربية المدعومة بالترميز الخاص بإظهار نظامها الخطي، وما يتوافر فيها من أدوات بحث ومعالجة. ورابعها: علاقة بناء المدونة اللغوية العربية الحاسوبية (أو أدوات معالجة اللغة

العربية) بطرائق الإفادة منها في البحث اللغوي. وخامسها: التصاحب اللغوي الفيرثي (نسبة إلى العالم فيرث، واضع نظرية التصاحب collocation theory في عام ١٩٥٧؛ انظر فيرث Firth ١٩٥٧؛ وروحاني Rouhani ١٩٩٤) وتحليل المتتابعات اللفظية المباشرة، والتقارب الدلالي، والعلاقة بين المواد المعجمية والقواعد اللغوية، والتفريق بين المدونة بوصفها نظرية والمدونة بوصفها منهجًا تطبيقيًا، وتطوير سنكلير Sinclair (١٩٩١، ٢٠٠٤) لهذه المفاهيم بما يتوافر في المعالجات الآلية، وترسيخ جريس Gries (٢٠٠٣، ٢٠٠٨، ٢٠٠٩، ٢٠١٠) لمفاهيم التصاحب نوعياً وكمياً من جهة التلازم القريب والبعيد، ومن جهة التلازم النحوي colligation، ومن جهة التلازم المعجمي النحوي collostruction والمجموعات الدلالية semantic sets؛ كل هذا عن طريق أهم الرزم الإحصائية المعمول بها في علم لغة المدونات (قارنها بـ لسانيات المدونات الإحصائية بـ «آر» statistical corpus linguistics with R عند جريس Gries ٢٠٠٩، ٢٠١٠). أما آخرها فيتعلق بالممكن والمحتمل والمأمول في الموضوعات الخمسة السابق ذكرها، مع إلحاق هذا الموضوع بجملة من الموضوعات البحثية المقترحة.

٥. ١. البحث اللغوي المعتمد على المدونة corpus-based والموجه بالمدونة corpus-driven

ثمة منهجان أساسيان في البحث اللغوي ومادته النصوص الرقمية (توجنيني بونيللي Tognini-Bonelli 2000)، فالأول -كما في العنوان الفرعي- ينطلق من قضية لغوية محددة بافتراضات وفرضيات وأسئلة بحثية في ذهن الباحث اللغوي، ويقوم بتفسيرها وتحليلها بالاعتماد على النصوص الرقمية في المدونة [٤٦]. أما المنهج الآخر فهو النظر إلى المدونة وتدوين ما يُمكن ملاحظته من مسائل لغوية صرفية وتركيبة ودلالية وسياقية، والتوجه من المدونة إلى صياغة فرضيات البحث وأسئلته.

٥. ٢. أنواع المدونات العربية الحاسوبية

تتنوع المدونات اللغوية الحاسوبية بتنوع النوع، والغرض، والعدد، والتصميم[٤٧]، والذي يهم في هذا السياق تلك الأنواع المتعلقة بالعدد، على الرغم من أهمية بقية الأنواع. وتُعزى أهمية العدد إلى كونها الأكثر شمولية لمستويات اللغة؛ بخلاف النوع والغرض اللذين تنحسر أبحاثهما على مستويات لغوية محددة، وأجناس كتابية معينة، وبخلاف التصميم الذي ينحصر اهتمامه على مجال تطوير تصميم المدونة ومعالجتها من حيث أنواع التوسيم والتحشية. ولا يعني هذا التصنيف عدم وجود مدونة حاسوبية عربية تتنوع في النوع والغرض والتصميم، فالمدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية King Abdulaziz City for Sciences and Technology's Arabic Corpus KACSTAC)) قد تطورت في نسختها الجديدة (مختلفة عن النسخة القديمة www.ksacstac.org؛ انظر الثبيتي al-Thubaity ٢٠١٤) من حيث توفّر خصائص بحثية دقيقة للكشافات السياقية والتكرارات والمتتابعات، والتي ستُمكّن من تحديد أسئلة وفرضيات لغوية أكثر دقة إن أحسن الباحث استعمالها. وعلى صعيد أنواع المدونات العربية من حيث العدد، فالنوع الأول يتضمن نصوص اللغة العربية المكتوبة والمنطوقة أو أحدهما، والـمُدخلة حاسوبيًا مع توفر محرك بحث search engine شبكي فيها من أجل البحث عن كلمة محددة في النصوص المدخلة فيها حاسوبيا، وإمكانية إظهار عدد مرات تكرارها، وإتاحة إخراج كشافاتها السياقية، كله من أجل كشف السلوك البيئي السياقي اللغوي الطبيعي لتلك الكلمة وما يرد قبلها وبعدها من وحدات معجمية نظمية syntagmatic lexical units تُعرف بمصطلح (المتصاحبات اللفظية collocations). والنوع الثاني يتضمن مدونتين منفصلتين؛ الأولى: عربية، والثانية: لغة غير العربية، ويُعمل على جمعها معًا وفق عدة معايير هي على النحو الآتي: تقابل الأوعية genres والمجالات domains من حيث النوع والحجم، وتوافق أزمنة إنتاج النصوص الحية في كلتا

اللغتين، وتقارب خيارات محرك البحث فيها (انظر ميكنري McEnery 2003: ٤٥٠؛ ميكنري وشياو McEnery and Xiao 2007: ٢٠). أما النوع الثالث من حيث العدد، فيُعرف بالمدونة المتوازية parallel corpus، وهذا النوع من المدونات ما زال في بداية انطلاقته بين العربية والإنجليزية (انظر العجمي al-Ajmi ٢٠٠٣؛ وانظر حول بداية معالجة المدونة المتوازية في: بياو Piao ٢٠٠٠، ٢٠٠٢، وبيكر وآخرين Baker et al ٢٠٠٦: ٩ فيما يتعلق بطريقة معالجة المحاذاة آليا). والفرق بين المدونة اللغوية الحاسوبية المتقابلة وتلك الموازية تكمن في أن نصوص المدونة الأولى من اللغتين لا تكون نتيجة لعمل ترجمي مسبّقا، بينما في نصوص الثانية يكون توازي النصوص مشروطًا بناتج الترجمة الفعلية الواقعية، وتُدخَل هذه النصوص آليًا في قوائم تمثل كل مُدخل تركيبي أو معجمي من اللغة المترجم منها إلى ذلك المدخل التركيبي والمعجمي في اللغة المترجم إليها، والذي يُعرف بالمحاذاة الآلية automatic alignment بين نصوص اللغة المصدر وترجماتها في اللغة الهدف[٤٨].

وبالنسبة للمدونات العربية الحاسوبية، يرى الباحث أن أكبر المدونات العربية الأحادية -كما أُشير إلى ذلك سابقًا- تتمثل في المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية KACSTAC، ومدونات العربية الشبكية Arabic Internet Corpora، ومدونة أرابيكوربوس [٤٩]arabiCorpus. ويتوفر في هذه المدونات محرك استعلام query engine شبكي يقوم من خلاله الباحث بكتابة كلمة ما والبحث عن استعمالاتها في النصوص العربية المحوسبة. وثمة في المقابل أدوات أخرى لمعالجة النصوص العربية الرقمية، ومختلفة عن تلك المتضمنة في المدونات الثلاث المذكورة آنفًا، وهي أدوات برمجية مستقلة عن المدعومة في مواقع تلك المدونات الثلاث، حيث تقوم بمعالجة النصوص العربية المحفوظة في امتدادات عديدة على غرار text و csv، و xml، و rtf وغيرها. ومن أهمها قبولا للتطبيق التحليلي في أكثر الأدوات ذكاءً حتى الآن (غواص

ACL أولًا ثم سكتش إنجن Sketch Engine). ويُرى النظر في أدوات معالجة العربية من أجل المقارنة بينها من حيث التصميم، والإتاحية، وقابلية القراءة للنصوص العربية، ونظام الاستعلام، ومؤشرات التصاحب، والنتائج الإحصائية ومقارناتها، وهي محور موضوع المبحث الآتي.

٥. ٣. أدوات معالجة النصوص العربية

تنقسم أدوات معالجة نصوص العربية إلى قسمين، أولهما: برامج تطبيقية حاسوبية حيث توفر للباحثين إمكانية تحميل نصوص عربية على امتداد ملف text أو امتداد ما يسمى بالقيم المفصولة بفواصل comma separated value csv، والتي تحول آليا في برنامج إكسل، وتوفر أيضا إمكانية البحث عن الكلمة العربية، شريطة أن تكون هذه الأدوات المخصصة لمعالجة اللغة العربية مبرمجة مسبقا من أجل دعم فك شفرات أشكال الكتابة العربية الإملائية بالترميز العربي الحاسوبي إلى أشكاله الكتابية الإملائية، أي: تكون مدعومة بأحد الترميزات الآتية: UTF-8، أو UTF-16 (الثبيتي وآخرون al-Thubaity et al 2013). ومن أنواع هذه الأدوات الداعمة لقراءة النصوص العربية: مونوكونك MonoConc[٥٠] والفهرسة الاستدراكية وأسلوب بناء الاسترجاع بلغة التوصيف الموسعة XML Aware Indexing and Retrieval Architecture Xaira[٥١] والكلمة الدالة في السياق KWIC (keyword in context) [٥٢] وأكونكورد aConCorde[٥٣] وأدوات وورد سميث WordSmith Tools[٥٤] وغواص Ghaww□□[٥٥]. ويتطلب من الباحث اللغوي عند استعمال هذه الأدوات برنامج الجافا Java من أجل تشغيل هذه الأدوات. أما ثانيهما: فهي أداة متوافرة على الشبكة العنكبوتية، وتُعرف باسم Sketch Engine[٥٦]، وتعد أداة معالجة للنصوص اللغوية (ومنها العربية) (تحدث الثبيتي عن بعض هذه الأدوات في المبحث الثالث من هذا الكتاب).

ويُتيح هذان النوعان من الأدوات نظام الاستعلام النصي عن الكلمات والعبارات وتكراراتها Frequencies في النصوص العربية واستخراج كشافاتها السياقية concordance lines. وكل أداة من هذه الأدوات تختلف عن الأخرى من حيث ما تتيحه من أنظمة استعلامية query systems، أي أن كل أداة من هذه الأدوات تتضمن أنظمة تُمكن المستخدم من تحديد النتائج المطلوب إظهارها ضمن قوائم على شاشة الحاسب، وتتوافق هذه الأدوات في بعضها، وتختلف في بعضها الآخر، وتتميز بعضها بأنظمة جديدة ليست متوفرة في الأخريات (انظر الجدول ٣٦).

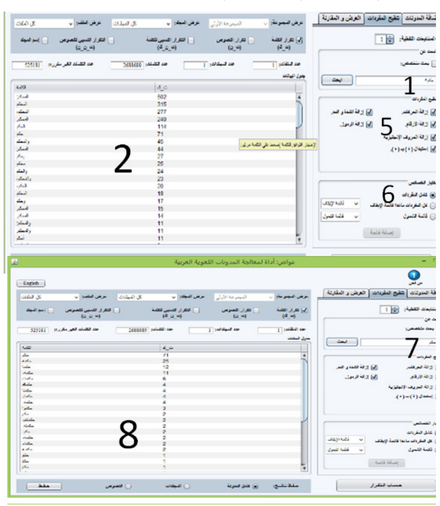
الأدوات	Mono Conc	Xiara	KWIC	aConCorde	WordSmith Tools	غواص	Sketch Engine
المصمَم	مايكل بارلو Michael Barlow	BNC	ساتورا تسوكاموتو Satoru Tsukamoto	أندرو روبرتس Andrew Roberts	مايك سكوت Mike Scott	الثبيتي وآخرون	Lexical Computing Ltd.
الإتاحية	متاح بالشراء	متاح في موقع BNC	متاح بالمجان	متاح بالمجان	متاح بالاشتراك	متاح بالمجان	متاح لتجريب مليون كلمة، ثم بالاشتراك
قابلية القراءة الآلية Machine-Readability	ترتيب الكلمات في الجملة ليس دقيقا في نتائج البحث عن كلمة	الترتيب دقيق	الترتيب دقيق	الترتيب دقيق	الترتيب دقيق	الترتيب دقيق	الترتيب دقيق
نظام الاستعلام Query system	متوفر بدون خيارات متقدمة	دقيق ويحتاج لسرعة عالية ويتطلب الكثير من الوقت	متوفر بدون خيارات متقدمة	متوفر ويتيح البحث عن الكلمة المفتاح، أو العبارة، أو باستعمال wildcard بالرمزين (*) أو (؟)	متوفر بدون خيارات متقدمة	متوفر مع خيارات متقدمة: استعمال wildcard وإمكانية تنقيح النص من الهمزة والتاء المربوطة والأرقام والرمز الأجنبية وغيرها	متوفر مع خيارات متقدمة باستعمال wildcard بالرمز (*) ونظام استعمال phrase و Corpus Query Language

مؤشرات التصاحب Collocation indicators	متوفرة بدون دقة	متوفرة	غير متوفرة	متوفرة بدون دقة	متوفرة بدون دقة	متوفرة ودقيقة مع تحديد مواقع امتداد التصاحب إلى مدى ١٥ كلمة	متوفرة، لكن مع عدم إمكانية تحديد مواقع امتداد التصاحب
الرُّزمة الإحصائية الخاصة بالمدونة Corpus-Based Statistics Package	غير مدعومة	مدعومة بـ Mutual Infor- mation و z-score	غير مدعومة	غير مدعومة	مدعومة بـ: Chi-Square Likelihood logDice	مدعومة بالرزم الآتية Chi-Square Weird Coefficients Mutual Information Likelihood t-score z-score LogDice	مدعومة بالرزم الآتية t-score Mutual Information logDice

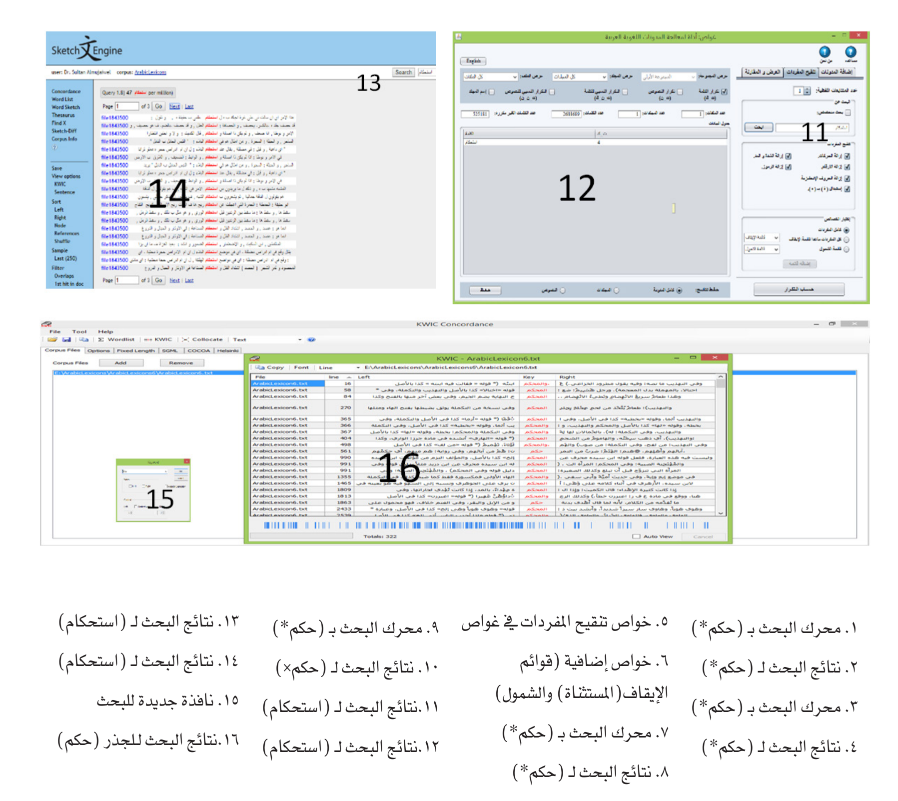
الجدول (٣٦) أدوات معالجة نصوص اللغة العربية وخصائصها

بالنظر إلى هذا الجدول، نلحظ أن الإتاحية وقابلية القراءة تبدو واضحة، بخلاف نظام الاستعلام، ومؤشرات التصاحب، والرزمة الإحصائية الخاصة بالمدونة. فنظام الاستعلام يراد به ما يمكن توافره من خيارات بحث تتيح للمستخدم إمكانات تتباين من حيث العموم (التطابق الكلي) والدقة (التطابق الكلي والجزئي ونوع الكلمة)؛ فالأدوات MonoConc، و KWIC، و WordSmith Tools لا تتضمن إمكانية البحث بطريقة المطابقة الجزئية باستعمال خاصية wildcard، أي: باستعمال الرمز (*) أو الرمز (؟). ونجد هذه الخاصية في الأداتين غواص و Sketch Engine، وتمكّن هذه الخاصية البحث عن التطابق الجزئي، فعلى سبيل المثال: لو أدخلنا (حكم*) فإن نتائج البحث ستظهر جميع المتطابقات للكمة التي تبتدئ بالجذر (حكم) وجميع وحداتها الصرفيمية السابقة المقيدة بها، ومن أمثلة النتائج: أحكم/استحكم/ المحكم/أراحكم/إلخ. (الكلمة: أراحكم ليست من جذر /ح/، /ك/، /م/ إلا أن طريقة المعالجة بالرمز × تظهر أي تطابق جزئي شكلي). أما لو أدخلنا (*حكم) فإن نتائج البحث ستظهر جميع المتطابقات الجزئية للجذر نفسه مع جميع وحداتها الصرفيمية

اللاحقة المقيدة بها، ومن أمثلة النتائج: حكمة/ حكمك/ حكمنا/ حكمهم/ إلخ. (انظر إلى الأرقام ١، و٢، و٧، و٨ في واجهة غواص، والأرقام ٣، و٤، و٩، و١٠ في واجهة Sketch Engine في الشكل ١٤) ويُرى هنا أن خاصية المحرف البديل wildcard character لا تضمن استخراج المتطابقات للوحدات الصرفية المقيدة الداخلة، فكلمة (استحكام) مثلا تحتاج إلى أن يُبحث عنها لوحدها، وكذا الحال لكل كلمة اشتقاقية طابعها النسقي الصرفيمي غير سلسلي non-concatenative (انظر إلى الأرقام ١١، و١٢، و١٣، و١٤ في الشكل ١٤). وما يلفت الانتباه في هذا السياق هو أن أداة KWIC Concordance برغم عدم توفر هذا الخاصية فيها، إلا أنه بمجرد كتابة الجذر (حكم) في خيار KWIC، فإن جميع النتائج تتضمن متطابقات هذه الكلمة الكلية والجزئية ذات النسق الصرفي السلسلي فقط (انظر إلى الرقمين ١٥ و١٦ الشكل ١). أما الرمز الثاني (؟) المتعلق بخاصية wildcard فإن وضعه قبل الكلمة المراد البحث عنها أو بعدها يعين على استخراج تطابقات هذه الكلمة الجزئية بإظهار حرف واحد فقط، على سبيل المثال: (حكم؟) = حكمة/حكمت/حكمك/ إلخ. و(؟حكم) = أحكم/يحكم/نحكم/ إلخ.







١. محرك البحث بـ (حكم*)
٢. نتائج البحث لـ (حكم*)
٣. محرك البحث بـ (حكم*)
٤. نتائج البحث لـ (حكم*)
٥. خواص تنقيح المفردات في غواص
٦. خواص إضافية (قوائم) الإيقاب (المستثناة) والشمول
٧. محرك البحث بـ (حكم*)
٨. نتائج البحث لـ (حكم*)
٩. محرك البحث بـ (حكم*)
١٠. نتائج البحث لـ (حكم×)
١١.نتائج البحث لـ (استحكم)
١٢.نتائج البحث لـ (استحكم)
١٣. نتائج البحث لـ (استحكم)
١٤. نتائج البحث لـ (استحكم)
١٥. نافذة جديدة للبحث
١٦.نتائج البحث للجذر (حكم)

الشكل ١٤: واجهة غواص وSketch Engine وKWIC Concordance ونتائج بحث كلمة (حكم)

أما مؤشرات التصاحب collocational indicators فهي عملية معقدة، ويرمز لها بـ N-gram، وتشير إلى المتتابعات اللفظية السابقة أو اللاحقة، والتي تمتد من ١ إلى ١٥، وتنطلق من المادة العنقودية Nodal item التي يُبحث عنها في الأصل بوساطة أدوات التحليل المذكورة آنفاً (انظر: ٢. ٥.)، وتظهر المتصاحبات (أو المتتابعات) اللفظية للمادة العنقودية على شكل كشافات سياقية concordance lines (الشكل ١٥) وحساب التوافق لكلمة معجم في مدونة فرعية للمعاجم العربية ArabicLexicon6).



الشكل (١٥) الكشافات السياقية للكلمة الأساس (معجم) بامتداد ٥ متتابعات لفظية سابقة ولاحقة

ويُرى أن حساب توافق الكلمة النوعية type والكلمة الفعلية token تختلف بين غواص وSketch Engine، فالأول يختلف عن الثاني في الحساب، ولزامًا على الباحث أن يُدرك سبب الاختلاف، والعائد إلى دقة غواص في تحليل مفهومي الكلمة النوعية والكلمة الفعلية والمسافات بينهما التي عادة ما تكون محشوة بعلامات الترقيم، وعليه لو وضعنا مدونة عربية في صيغة text، وقمنا بتحميلها في أداة غواص والأداة الشبكية Sketch Engine، فإن حساب التطابق في الثاني سيكون أعلى نظرًا لاعتباره المسافات المملوءة بعلامات الترقيم وغيرها من الرموز. ولكن يتميز Sketch Engine بإمكانية معالجة النصوص العربية المتوفرة على المواقع الشبكية من خلال خاصية WebBootCat (انظر http://www.sketchengine.co.uk/documentation/wiki/Website/Features#WebBootCat).

والرزم الإحصائية الخاصة بالمدونة مجموعة من القياسات الإحصائية statistical measurements التي تقدم دلالات رقمية يكون منها القبول وعدم القبول في قياس قوة الارتباط بين الكلمات المتصاحبة (كلمة بصحبة كلمة) أو المتجاورة (كلمة بجوار كلمة يفصل بينهما كلمة أو كلمتين أو أكثر). ويُرى هنا تلخيص وظائف أهم هذه الإحصاءات المتعلقة بالتحليل الآلي للنصوص العربية، والتي تفيد الباحث اللغوي بالدرجة الأولى، وهي على النحو الآتي:

أولاً: ما يتعلق بالتحليل بين مجلدين (مدونتين)، ومن أهمها: مربع كاي Chi-Square (x2) حيث تفيد في قياس توزيع تكرارات النصوص أو الكلمات بين مجلدين بحساب التكرارات الملحوظة observed frequencies مع التكرارات المتوقعة expected frequencies (والأخيرة تُستخلص آلياً من الأولى) من أجل دعم الفرضية الصفرية null hypothesis. وأداة غواص تقوم بحساب مربع كاي بشكل آلي، ولا يُفضل استعماله لمعرفة قياس تكرار الكلمات في المجلد الواحد. والسؤال المهم هنا: كيف يُمكن للباحث اللغوي الذي سيحلل نصا رقميا باستعمال أداة غواص أو Sketch Engine أن يستفيد من قراءة الأرقام الإحصائية التي تقدمها هاتان الأداتان فيما يتعلق بهذا القياس؟ وكيف لهما أن تدعما أو ترفضا الفرضية الصفرية التي يصوغها الباحث؟ عند مقارنة مدونتين (مجموعتين) من حيث النظر، على سبيل المثال، إلى تكرارات كلمة ما، فإنه كلما كانت الفرضية الصفرية null hypothesis محددة فإنها لن تتوافق إلا بطبيعة المدونة المحللة، والمدخلة في أدوات التحليل. بمعنى آخر: لو قمنا بجمع نصوص المعاجم العربية كلها قديما وحديثا (٢٤ معجمًا)، واستخلصنا فرضية لغوية صفرية مفادها: أن اللواصق التصريفية inflectional suffixes أكثر من اللواصق الاشتقاقية derivational suffixes فيها بخلاف مدونة الصحف السعودية حيث اللواصق الاشتقاقية فيها أكثر من اللواصق التصريفية، فإن العينة هنا يجب أن تتوافق مع أركان الفرضية الصفرية، وهي أن الفروق بين

هذين النوعين من اللواصق واقعٌ طبيعيًا بين المدونتين بمحض الصدفة، وإن لم يكن وقوعهما بمحض الصدفة by chance، فإن ذلك يعني أن الفرضية لا يُمكن قبولها. ولكن كيف يُمكن قراءة الأرقام المتعلقة بهذا القياس؟ كلما اختلف حجم كلٍّ من المجموعتين (كل مجموعة تتضمن ملفًا لنصوص رقمية عربية)، اختلفت القياسات المتعلقة بمربع كاي؛ وعليه تختلف الفرضية الصفرية التي تُدعم بقيمة كل من التكرارات الملحوظة observed الواقعية والتكرارات المتوقعة expected، والتي تُستخلص بشكل آلي وفقاً لمعطيات المعادلة الآتية: (E-O)2/E □ وحساب هذه المعادلة ببساطة هو استخلاص نتيجة التكرار المتوقع expected frequency عن طريق جمع معطيات التكرارات الملحوظة (الواقعية فعليًا) من كل مجلد مستقل، واستخلاص نتائج جمعها عمودياً وأفقياً (الجدول ٣٧).

	تكرار اللواصق الاشتقاقية للفعل (ذهب) الملحوظة	تكرار اللواصق التصريفية للفعل (ذهب) الملحوظة	المجموع
مدونة المعاجم العربية	٦	١١	١٧
مدونة الصحف السعودية	١٢	١٠	٢٢
المجموع	١٨	٢١	٣٩

الجدول (٣٧) المرحلة الأولى لطريقة حساب التكرارات المتوقعة

في المرحلة الثانية، تُضرب كل قيمة لمجموع كل من التكرارات الملحوظة، ثم تقسم على مجموع نتائج التكرارات العمودية والأفقية: 7.8 =(39)/18*17 و10.1 =(39)/18*22، و9.1 =(39)/21*17، و11.8 =(39)/21*22 (الجدول ٣٨).

ثم تقسم نواتج هذه القيم باستعمال المعادلة الآتية: (O-E)2/E من أجل استخلاص أربع قيم (انظر الخانات المضللة في الجدول ٣٩)، ثم تجمع هذه

القيم لاستخلاص مربع كاي بشكل نهائي (مجموع حاصل (O-E)2/E من كل خانة: 1.51 =0.415+0.396+0.396+0.303 مربع كاي).

	تكرار اللواصق الاشتقاقية للفعل (ذهب) الملحوظة O	تكرار اللواصق التصريفية للفعل (ذهب) الملحوظة O	تكرار اللواصق الاشتقاقية للفعل (ذهب) المتوقعة E	تكرار اللواصق التصريفية للفعل (ذهب) المتوقعة E	المجموع
مدونة المعاجم العربية	٦	١١	٧,٨	٩,١	١٧
مدونة الصحف السعودية	١٢	١٠	١٠,١	١١,٩	٢٢
المجموع	١٨	٢١	١٨	٢١	٣٩

الجدول (٣٨) التكرارات الملحوظة observed frequencies الفعلية والتكرارات المتوقعة الآلية expected frequencies

	مجموع طرح قيمة الملحوظة على المتوقعة (O-E) للواصق الاشتقاقية	(O-E)2/E	مجموع طرح قيمة الملحوظة على المتوقعة (O-E) للواصق التصريفية	(O-E)2/E
مدونة المعاجم العربية	١,٨	٠,٤١٥	١,٩	٠,٣٩٦
مدونة الصحف السعودية	١,٩	٠,٣٩٦	١,٩	٠,٣٠٣

الجدول (٣٩) القيم المحسوبة calculated values للتكرارات الملحوظة والمتوقعة

هذه القيمة المحسوبة calculated value تعد أقل من القيمة الواقعية critical value؛ أي عدد الأسطر والأعمدة من المتغيرات في حساب مربع كاي في المربع في الجدول (٣٩). وهنا فإن كل عملية حسابية للمتغيرات (أو ما يُصطلح عليه

باسم درجة التكرار degree of frequency، والذي يمثل قيمة ٣ في المثال المذكور آنفا؛ أي: الخانة الأفقية الأولى الخاصة بالفرضية مع خانتي القيم الأفقية للمدونتين كما هو في الجدول ٣٩) تتضمن القيمة المحسوبة والقيمة الواقعية، وكل فرضية لغوية تُحسب بوساطة مربع كاي تكون نتائجها وفق معيارين؛ الأول: تُرفض الفرضية اللغوية الصفرية null linguistic hypothesis عندما تكون القيمة المحسوبة calculated value أعلى من ٠,٠٥ والثاني: تُقبل الفرضية اللغوية الصفرية عندما تكون القيمة المحسوبة أقل من ٠,٠٥، وعليه فإن القيمة النهائية لمربع كاي ١,٥١ تُعد أعلى من القيم ذات الدلالات الإحصائية: ٠,٠٥، أو ٠,٠١، أو ٠,٠٠١، وعليه إن الفرضية التي صُغناها (اللواصق الاشتقاقية والتصريفية في المدونتين مختلفة) غير صحيحة.

ثانيًا: فيما يتعلق بتحليل الكلمة المبحوث عنها ومتصاحباتها (الكلمة السابقة أو اللاحقة مباشرة) أو متجاوراتها (الكلمة السابقة أو اللاحقة غير المباشرة؛ أي: في الموضع الثاني أو الثالث أو الرابع أو الخامس). ومن أهم الإحصاءات التي تفيد دلالة التصاحب هنا هي: المعلومات المتبادلة Mutual Information (MI)، وقياس-ت T-Score، وقياس-ز Z-Score، ومعامل دايس Dice واللوج دايس logDice.

فالمعلومات المتبادلة قد وضحها تشرتش وهانكز Church and Hanks (١٩٩٠) وأواكس Oakes (١٩٩٨) على أنها تفيد بالكشف عن احتمالات تكرار وقوع كلمتين يكونان متصاحبين معًا مرة، ووقوعهما منفصلتين. ومعادلة هذه الإحصائية هي:$\log2((P(x,y)/P(x)P(y))$ حيث إن P هو عدد تكرار الكلمة، أما x و y فهما المتصاحبان اللذان يُراد اختبارهما. ولو قمنا على سبيل المثال باختبار تكرار كلمة (عاصفة أو عاصف) مع الصفتين (شديدة) و(قاشرة)، فإن المعطيات الإحصائية ستكون على النحو الآتي: عدد كلمات مدونة المعاجم العربية ٢٠,٤٣٢,٢١٢ كلمة، وعدد تكرار كلمة (عاصفة أو عاصف) فيها ١٢٠

مرة، وعدد تكرار (عاصفة قاشرة) ٤ مرات، وعدد تكرار (عاصفة شديدة) ٥ مرات، وعدد تكرار الصفة (قاشرة) لوحدها ١٢ مرة، وعدد تكرار الصفة (شديدة) لوحدها ١٥١٩ مرة. ولقياس كل متصاحب collocate من الصفتين (أي: شديدة وقاشرة) على حدة مع المادة العنقودية nodal item (أي: عاصفة أو عاصف)، فإن المعادلة تتمثل على النحو الآتي:

أولا: معادلة المعلومات المتبادلة للتصاحب «عاصفة أو عاصف شديدة» وناتجهما هي:

log2((5 * 20,432,212)/(120 * 1519))= 9.13

ثانيًا: معادلة المعلومات المتبادلة للتصاحب «عاصف أو عاصفة قاشرة» وناتجهما هي:

log2 ((4 * 20,432,212)/(120 * 12))= 15.7

بنتائج كل معادلة، نلحظ أن قوة التصاحب بين المادة العنقودية والصفة (قاشرة) أعلى من تلك الواقعة بينها والصفة (شديدة)؛ وقوة التصاحب هنا يُشير إلى أن التزام الصفة الأولى أكثر دلالة من حيث الاستعمال المنسجم من الصفة الثانية. وفيما يتعلق بدلالات المعلومات المتبادلة إحصائيا، فإن الناتج الذي يكون أعلى من ٣، يكون مقبولاً من حيث الدلالة الإحصائية، وإن كان أقل فإن دلالة التصاحب تكون حينها غير مقبولة.

وقياس-ت t-score يتشابه مع المعلومات المتبادلة غير أنه يقوم بإظهار مقاييس التشتت لاحتمالات تكرارات التطابق للمادة العنقودية nodal item ومصاحبها collocate بالاستناد إلى عدد الكلمات في المدى الواردتين فيه (سكوت Scott ٢٠١٠، وانظر هانستون Hunston ٢٠٠١، ٢٠٠٢؛ وبرايس Price ٢٠١٣). ويتكون هذا القياس من المعادلة الآتية: ((x/n)-x)/√j حيث إن n يعبر عن العدد الكلي للكلمات في المدونة، و j يعبر عن حاصل ضرب التكرار المشترك بين

المادة العنقودية ومصاحبتها؛ حيث إن x يشير ببساطة إلى F1*F2 (أي: ضرب عدد تكرار الكلمة الأولى مع عدد تكرار الكلمة الثانية المصاحبة للكلمة الأولى). ويُستفاد من هذه العملية في تحليل الكلمات ذات المعاني المتعددة polysems. والناتج الآلي الذي يُستخرج بهذه المعادلة يجب أن يكون من ٢ فما فوق من أجل ضمان قياس إحصائي ذي دلالة قوية.

أما قياس-ز z-score، فقد شرحه بيري-روجي Berry-Rogghe (١٩٧٣: ١٣١) بالمعادلة الآتية:

Z: العدد الكلي للكلمات في المدونة

A: المادة العنقودية وعدد مرات تكراراها

B: المصاحب للمادة العنقودية وعدد مرات تكرارها

K: عدد مرات تكرار A و B معًا

S: المدى وعدد الوحدات المعجمية قبل المادة العنقودية وبعدها

تقوم الأداة المدعومة بهذه الرزمة الإحصائية باستخراج قيمة z-score بين المادة العنقودية والمتصاحب أياً كان موقعه (مدى خمس كلمات5-word span)، وكل متصاحب لفظي مع المادة العنقودية تكون قيمته أعلى من ٣ يُعد دالًا إحصائيا بشكل مقبول.

أما آخر وأهم قياس إحصائي بين المادة العنقودية ومصاحبها فهو الدايس Dice (كيلجارف وآخرونKilgariff et al 2004)، وأساس معادلته هو /2fAB (fA+fB)) حيث إن fAB يدل على تكرار الوحدة المعجمية الهدف مع المتصاحب المعني بالتحليل، و fA+fB يدل على حاصل جمع تكرار المادة العنقودية لوحدها مع تكرار المتصاحب المعني بالتحليل. ومشكلة هذا القياس هو أنه يقدم قيمًا صغيرة جدًا (أقل من ٠,٠)، أما معامل اللوج دايس logDice فيضاف إلى

معادلة الدايس الآتي:logD +14 2fAB/(fA+fB) . فلو نظرنا إلى المثال الذي ذكرناه آنفا حول المادة العنقودية (عاصف أو عاصفة) مع الصفتين اللتين تُصاحبها في مدونة المعاجم العربية: (شديدة وقاشرة)، فإن الحسبة الآلية تكون على النحو الآتي:

تكرار الصفة المتصاحبة (شديدة) للكلمة الأساس (عاصف أو عاصفة) وفقًا لتكرارات كل واحدة على حدة مع تكرارهما معًا وبناء على الدايس ومعادلته 2fAB/(fA+fB) تكون النتيجة: ٥/(١٢٠+١٥١٩)= ٠,٠٠٣٠٥ وقيمة اللوج دايس ١١,٥، أما تكرار الصفة المتصاحبة (قاشرة) للكلمة الأساس (عاصف أو عاصفة) وفقا لتكرارات كل واحدة على حدة مع تكرارهما معًا وبناء على معادلة الدايس 2fAB/(fA+fB)تكون النتيجة 0.0303 =(120+12)/4. وتكون نتتيجة اللوج دايس ١٢,٥. وبناء على النتيجة الأولى والثانية نجد أن الثانية الأعلى من الأولى ذات قيمة عليا، وينظر للأعلى أو الأقل بحسب ما يريده الباحث من الفرضية من حيث قوة التصاحب من عدمه[٥٧].

٥. ٤. علاقة بناء المدونة اللغوية العربية الحاسوبية وأدوات التحليل بطرائق البحث اللغوي

على كل باحث في بادئ ذي بدء أن ينطلق من فرضياته اللغوية ثم يُوجد المدونة اللغوية العربية الحاسوبية أو الأدوات اللغوية أو بهما معا، والتي يُمكن لها أن تجيب عن تلك الأسئلة، أو أن ينظر إلى خصائص المدونة اللغوية العربية الحاسوبية من أجل أن ينطلق من تلك الخصائص التي تُكيف أصلا أسئلة البحث اللغوي. ومن كلا الجهتين ينطلق الحدس اللغوي العميق أولًا، ثم الآلة ثانيًا، ثم بهما معا شريطة أن يكون إعمالهما إزاء بعض منهجيًا ومقبولًا في المحصلات النهائية للتجربة، ويكونان معًا برهانًا للفرضية اللغوية الصفرية المصوغة، والمدونة المحددة للاختبار. وهنا تكون إجابات الأسئلة التي ذكرناها

في بداية المبحث؛ ومفادها: هل من الممكن أن تُجيبنا المدونات عن كل أسئلة البحث اللغوي؟ وهل للبحث اللغوي المعتمِد على المدونات شروط منهجية؟ ولماذا؟ فجواب (هل) هنا هو أن كل مدونة لغوية حاسوبية يستحيل أن تجيبنا عن كل الأسئلة؛ لأن الأسئلة هنا تتحدد بماهية المدونة بحسب نوعها أو غرضها أو عددها أو تصميمها (ميكنري وهاردي McEnery and Hardie 2012: 27).

أما الشروط المنهجية فتتحدد بطريقة يُوفّقها الباحث اللغوي مع العلماء المختصين بمعالجة اللغة الطبيعية Natural Language Processing. وأما السؤال (لماذا؟) فهو أساس البحث اللغوي العام أو الخاص، أي: أساس المنهج المتبع بتتبع نوع المدونة، وتتبع نتائج الرزم الإحصائية، وتوافق تلك النتائج بالفرضيات اللغوية.

٥. ٥. تحليل التصاحب اللغوي

يجب تبسيط مفهوم التصاحب أولًا، وهو مفهوم مركب يمكن تقديمه على النحو الآتي[٥٨]:

أولاً: في عملية تحليل المعاني المركبة بمجموعة من المدخلات المعجمية المتصاحبة في النصوص، فالتصاحب collocation يُعد ارتباطًا تركيبيًا أوليا لوحدتين معجميتين معًا في سياق لغوي معين، مثل: يستقل/يركب/يستعمل القطار، وقد يخرج عن المتعارف عليه عند مزيد من الكشف عن مستويات النصوص اللغوية العربية الرقمية.

ثانياً: في عملية تحليل المعاني المركبة بمجموعة من المدخلات المعجمية المتلازمة في النصوص، فالتلازم colligation يُعدّ إيقاعا تركيبيًا إلزاميا لوحدتين معجميتين معًا في أي سياق، مثل: الفعل اللازم والفعل المتعدي وحروف الجر وما بعدها من أسماء معرفة بـ (أل) وغيرها من المتلازمات.

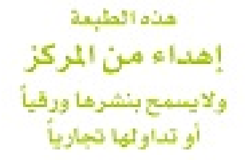
ثالثاً: في عملية تحليل المعاني المركبة نصًا بمجموعة من المدخلات المعجمية المتجاورة في النصوص، فالتجاور collostruction يعد إيقاعا تركيبيا متغيّرا لوحدات معجمية نظمية توليدية في سياقات عديدة؛ مثل: المكملات أو المتممات complements في توليد بقية الجمل الإسمية والفعلية الأساسية من الأحوال والصفات والتمييز وأدوات الربط conjunctions التي تولد مزيدًا من التوليد، والتلازمات النحوية التي تزيد عن أكثر من ثلاث كلمات؛ مثل: الفعل المتعدي، ولا النافية للجنس، والأفعال الناسخة، وأخوات إن واسمها وخبرها، وظن وأخواتها، والالتزامات المتتابعة بمزيد من التجاور النظمي، كل ذلك في سياق تجاوري نظمي يُمكن من خلاله تحليل قواعد التركيب construction grammar (جولدبيرج Goldberg 2009).

رابعًا: في عملية تحليل المعاني المركبة بمجموعة من المدخلات المعجمية المتباعدة في المدى span والمتقاربة في الدلالة (المجموعة الدلالية semantic set أو التقارب الدلالي semantic preference) يُعدُّ التضام هو المحك، وكثرة التعارف عليه تكشفه اللغة الطبيعية المحوسبة في المدونة الضخمة large-scale corpus؛ مثل: اصطادَ السمك من قارب أو زورق (لا يحدث في العادة اصطياد السمك من متن باخرة أو سفينة).

ودراسة هذه الأنواع قد تكون على مستوى الكلمات أو مستوى النصوص من جهة، وعلى مستوى التصاحب مع كلمة في نص وتوافق هذا التصاحب من عدمه في نص آخر من جهة أخرى. وأول من اهتم بنظرية التصاحب هو فيرث Firth (١٩٥٧)، وجاء من بعده سنكلير Sinclair (١٩٩١، ٢٠٠٤)، والفرق بينهما هو أن الأول قد اهتم بالمدى span بغض النظر عن الموضع position، واهتم الثاني بالموضع بغض النظر عن المدى. أما مفاهيم جريس Gries التي ذكرناها آنفًا فهي أكثر دقة من كلٍّ من فيرث وسنكلير.

وقوائم التصاحب اللفظي التي يمكن الحصول عليها بأدوات معالجة العربية قد لا تُجيبنا كثيرًا عن السؤال اللغوي العام: ما فائدة هذه القائمة من الناحية اللغوية؟ ولعل أعم إجابة يُمكن أن تبرهن عن فائدة واحدة لهذا القائمة هو أن بعض التكرارات المتعلقة بأدوات الربط conjunctions والحروف particles (حرف الجواب والنفي والشرط والتوكيد والتمني والصلة والترجي والنداء والأمر والجر والنهي والجزم والحروف المشبهة بليس والحروف المشبهة بالفعل إلخ)، وأسماء الإشارة والضمائر قد تعطينا ملمحًا عن نتائج هذه المدونة المحددة بحدود ماهيتها من حيث النوع والعدد والحجم، وهو ملمح قد لا يتجاوز فكرة طريقة التركيب والاستعمال التركيبي بين كل هذه الأدوات وما يرد بعدها وقبلها، وما تفيده في تطبيقات التحشية الإحالية anaphoric annotation من جهة، وأشكال التنوع variation الاستعمالي من وجهة نظر لغوية سوسيولوجية.

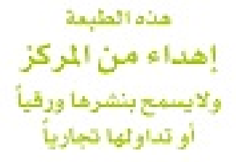
وثمة مناحٍ أخرى جديرة أوليًا بالاهتمام في مجال التحليل التصاحبي في أي مدونة، ومنها: قضية التصاحب collocation والتجاور collostruction. ولو أخذنا على سبيل المثال ما يوفره Sketch Engine في نظام لغة استعلام المدونة Corpus Query Language CQL (ياكوبيتشيك وآخرون Jakubíček et al، ٢٠١٠) والذي يُتيح البحث في المدونة عن طريق خواص الترميز [attribute=»value«]النوع صيغة الكلمة أو بالاعتماد على رموز التوسيم للغة العربية (مجموعة وسوم ستانفورد ومنى دياب لأقسام الكلام Stanford's Mona Diab's Part-Of-Speech Tagging Sets: ومنى دياب يُعرف بـ AMIRA وهو مطور من توسيم MADA؛ انظر: دياب Diab ٢٠٠٧، ٢٠٠٩ وانظر حول مدى MADA عند: حبش وآخرين Habash et al 2009). ولو أراد الباحث على سبيل المثال استعمال هذه الخاصية من أجل الكشف عن البحث على استعلامات الضِمّن within and containing queries، أو على استعلامات الاتحاد meet



and union queries فستُتيح الخواص الترميزية لكل مشغل على حدة فرصة استخراج بحث عالٍ من الدقة من حيث التوافق في نظام استعلاء CQL (انظر الشكل ١٦)، وهذا كله يُسمى بـ «مشغل الضمن أو الاتحاد withingand containing or meet and union operator». (الشكل ١٦).

الشكل (١٦) برنامج المستخدم للنظام: نافذة حوار نظام الاستعلاء في Sketch Engine

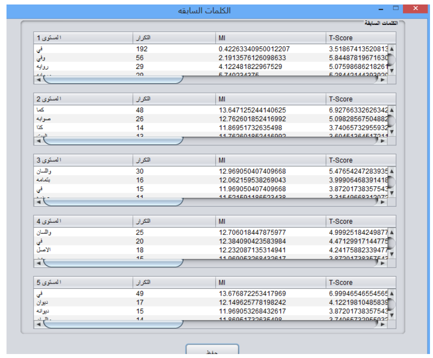
في الشكل (١٦) نلحظ أن المدونة المستعملة هي مدونة المعاجم العربية (٢٤ معجمًا عربيا قديما وحديثا). وبالنظر إلى هذا الشكل، نلحظ خاصية الاستعمال CQL التي تعتمد على أساس خواص الترميز لكل مشغل مذكور آنفًا. ومن الممكن للباحث اللغوي هنا أن يتصدى للحروف والأحوال وأدوات الربط والتعابير الاصطلاحية والجمل الفعلية، وذلك بنظام التوسيم الذي تقوم عليه Sketch Engine (ستانفورد STANFORD وأميرة AMIRA؛ انظر: دياب Diab 2009). وفيما يلي نشير إلى مثال تطبيقي لمشغل الضِمّن أو الاتحاد:

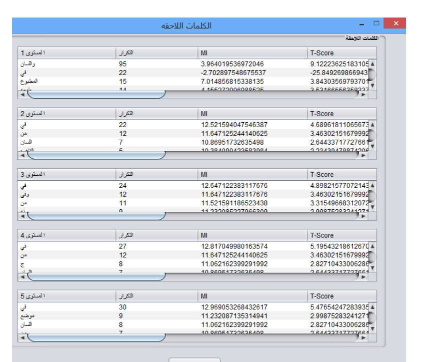


within	[tag=»PR.*»] within [tag=»V.*»] [tag!=»V.+»] [tag=»PR.*»]
الضِمْن	توسيم يُظهر الأفعال اللازمة في مدونة المعاجم العربية، ونتيجته: ٤٦٦,٤٢٣ من أصل ٢١٢,٤٣٢,٢٠ كلمة
Containing	
meet	[tag=»V.*»][tag!=»V.+»][tag=»PR.*»]
الاتحاد	توسيم يظهر الفعل العربي وامتداده في نصوص مدونة المعاجم العربية، :
Union	ونتيجته ٠٥٠,٤٢٢ من أصل ٢١٢,٤٣٢,٢٠ كلمة

وهذان الترميزان أنموذجان لطريقة البحث في Sketch Engine، وثمة العديد من الترميزات التي تتيح الكشف عن أنواع مواد الكلمة المعجمية lexical word-items (الجمل الفعلية والمتصاحبات الاسمية من الأعلام وأسماء المؤسسات والأفعال اللازمة والإضافة والتعابير الاصطلاحية والتضام الدلالي المنسجم؛ انظر إلى نظام توسيم أقسام الكلم للعربية الخاص بأميرة في: حبش ٢٠١٠: ١٧٥)، والتي يستطيع الباحث اللغوي أن يستعين بها من قبل المختصين بلسانيات المدونات.

وتُمكّن أداة غواص لتحليل مفاهيم التصاحب عن طريق المادة العنقودية nodal item وما يتبعها من متصاحبات مع إمكانية تقديم تفسيرات إحصائية لقوة التصاحب في مواضع positions أربع: الكلمة الأولى، والكلمة الثانية، الكلمة الثالثة، والرابعة، وهي مزية تُحسب لأداة غواص، وتتميز بها، ولا توفرها أداة Sketch Engine (الشكل ١٧).





الشكل (١٧) نتائج حساب التطابق من حساب موضع الكلمات السابقة واللاحقة
إحصائيا بـ MI وt-score وz-score و logDice

وبالنظر إلى هذا الشكل، سنلحظ أن التتابع السابق يعني: المتتابعات السابقة، والتتابع اللاحق يعني: المتتابعات اللاحقة، والمستوى يعني: الموضع position الذي يعكس موقع التتابع بوصفه المتتابع الأول (المستوى الأول)، أو المتتابع الثاني (المستوى الثاني)، أو المتتابع الثالث (المستوى الثالث)، أو المتتابع الرابع (المستوى الرابع). وبهذه المزية المعالجاتية، فإن للباحث اللغوي القدرة على كشف مواضع التتابع المباشر وغير المباشر مع قراءة إحصاءاتها المتعلقة وفق ذلك الموضع.

الاتجاهات البحثية الممكنة

لا يمكن حصر كل الاتجاهات الممكنة، ولكن ثمة ما قد يفتح للباحث اللغوي آفاقا ما إن يُعمل بها فإنها ستدفع بما سيشغله في الدرس اللساني. وتعد الاتجاهات البحثية اللغوية والترجمية والمعجمية بوساطة الاستناد إلى مدونة لغوية عربية حاسوبية[٥٩] أولية من جهة الواقع الحيوي للوحدات الجذعية وما يلتصق بها من لواصق اشتقاقية وتصريفية مستجدة، والتراكيب اللغوية وما يتغاير فيها من حيث الأهمية لمواد الكلمات المعجمية lexical word-items، والتقارب الدلالي

semantic preference، والكشف عن بيئات هذه الوحدات المعجمية الاستعمالية من حيث الاشتقاقات والتصاحب والدلالات اللغوية وغير اللغوية (اجتماعية، أو نفسية، أو أكاديمية، إلخ).

الاتجاهات البحثية المحتملة

من الاتجاهات البحثية المحتملة: إعادة النظر من جديد في مفاهيم التصاحب اللغوي، وتلك التقسيمات المذكورة آنفا (انظر ٥. ٥.) التي تحتاج أبحاثًا على مستوى المؤسسات والأفراد، وسيفرض هذا الاتجاه رسم الأبعاد الاستعمالية للوحدات المعجمية في المعجم الذهني العربي، والوصول لقواعد بيانات تتضمن صورًا وأشكالًا لتصاحب الألفاظ عن طريق تصاحبها الحر والمقيد لفظيًا، وتلازمها نحويا، وتجاورها تركيبيا، وتضامها دلاليا.

إن قواعد بيانية حاسوبية لهذه الأحوال التصاحبية الأربعة ستفتح آفاقًا لمعرفة لغوية عربية حية تتولد من عقول مستعملي اللغة، وتعيش بعيدة عن البعثرة من جهة الحدس، وأي لغة لها تاريخ استعمالي تاريخي طويل (كحال العربية) يتحتم فيها أن تُمثَّل في تلك القواعد البيانية في ثلاثة اتجاهات؛ الأول: تعاقبي diachronic والثاني: آني (تزامني) synchronic، والثالث: التوقيف compromise (اللغة التوقيفية compromise language؛ أو اللغة الوسيطة) [٦٠] التي تتوسط بين التغيرين الزمني والوصفي من جهة والتقليدي واللهجي من جهة ثانية والعمل المؤسسي على تحديد المستويات النموذجية standard levels من جهة ثالثة (انظر: مول Moll ٢٠٠٨-٢٠١٠، وانظر الحمزاوي ١٩٨٦)؛ كل ذلك من أجل السعي إلى التمثيل البياني المقبول وصفًا واستعمالًا نموذجياً للغة العربية.

الاتجاهات البحثية المأمولة

من المأمول مستقبلاً توسيع دوائر معالجة اللغة العربية بأدوات حاسوبية يُمَكن لها أن توفر للباحث اللغوي العربي مزيدًا من التحليل، وذلك عن طريق تطوير معالجات التوسيم النحوي part-of-speech tagging (أو التوسيم القواعدي grammatical tagging) واستحداث بدايات جادة للتوسيم الدلالي semantic tagging والإحالي anaphora والنبري prosodic، والتي ستُفيد الباحث اللغوي بمزيد من وقائع التحليل اللغوي لوقائع اللغة العربية الحية.

ومن المأمول أيضًا أن يكون هناك اهتمام بالمدونات المتقابلة comparable corpora والمدونة المتوازية parallel corpus، حيث تفيد الأولى في الكشف عن المتقابلات السياقية اللغوية الاستعمالية بين مدونتين تتوافقان في المعطيات الزمنية والوعائية والمجالاتية والموضعية والسياقية. وهذا المرحلة كلما زادت جدّتها زاد عونها للمختصين في معالجة اللغات الطبيعية بفهم أعمق لآليات الترجمات الآلية؛ أما الثانية فستفيد في معرفة مزيد من المقابلات الترجمية السياقية الطبيعية، والتي سيكون لها دور في فهم أيّ المقابلات من اللغة الهدف أولى استعمالا من حيث الأكثر تكرارًا من جهة السياق.

موضوعات مقترحة في البحث اللغوي الحاسوبي

يعتمد البحث اللغوي على تحليل معلميّن أساسين من وجهة نظر المعجميات التركيبية structural lexicology (انظر ليبكا Lipka 2002)، وهما: المعلم الضملغوي (تحليل اللغة بالنظر إلى مكوناتها الداخلية بنيويا من الصوت إلى الدلالة) والمعلم الفولغوي (تحليل اللغة بالنظر إلى مكوناتها الخارجية وجوديا من النفس فالتربية فالتعليم فالثقافة فالمعرفة وصولا إلى التكوين الأم: علم الوجود).

وستُقترح هنا موضوعات في البحث اللغوي الحاسوبي تكون مادتها اللغة الحية (أي مدونة لغوية حاسوبية محفوظة في امتداد txt أو csv)، وأدوات تحليل هذه المادة اللغوية (وأفضلها: غواص وSketch Engine).

وأذكر هنا بعض الموضوعات التي يُمكن أن تُقترح، ليس من أجل بحثها بل من أجل تقريب الصورة الأولية لمنطلقات البحث اللغوي في المدونات العربية الحاسوبية وتحديد نوع المدونة. ومن ثم فالفرضية تكون صياغتها بداية تحديد البنية الصورية thematic structure لأية ورقة بحثية لغوية معتمدة على المدونات العربية وأدوات تحليلها ومعالجتها آليا.

نوع التحليل	المدونة المقترحة	الفرضية المقترحة
أنواع التراكيب النحوية من حيث التصاحب والتلازم	معاصرة (صحف/ سرديات وروايات)	التراكيب المتنوعة بتنوع المجالات
أنواع التراكيب من حيث التجاور والتضام	معاصرة (صحف/ سرديات وروايات)	التراكيب المتنوعة بتنوع الأوعية
أنواع التراكيب الاسمية من حيث أنواع مواد الكلمات المعجمية lexical word-items	مقارنة بين مدونة تراثية ومدونة معاصرة	مواد الكلمات المعجمية وأنواعها في العربية
تحليل الوجود للمواد المحسوسة	الكتب التراثية المتعلقة بالرحلات	مسميات الموجودات والمحسوسات في الكتب العربية التراثية
تحليل التعبيرات الخاصة بدقائق المواد الثقافية	الكتب التراثية	أسماء أجزاء كل محسوس أو موجود على حدة

التغاير بين التماثل المعجمي isolexical والتماثل النصي isotexutal	الكتب التراثية/ أو الصحف العربية القُطرية	نسبة التطابق والتغاير بين الوحدات المعجمية ومجالاتها الموضوعية
الروايات	الرواية العربية أو المحددة عربيا من حيث الجغرافيا السياسية	الاتجاهات الاستعمالية للوحدات المعجمية
الكلمة في السياقات الشعرية	الشعر العربي القديم والحديث	التنوع في الدلالات الشعرية
تحليل توافق الثقافة واختلافها للوحدات المعجمية	المدونات المتقابلة	بين لغتين فيهما أثر التشابه أو الاختلاف بين الدوال اللغوية ومدلولاتها
التطابقات لجميع الأفعال الزائدة عن الجذر root والجذع stem	إعمال tokenization و lemmatization لأجل التصدي لها	الاشتقاقات القديمة والجديدة
السير	السير الذاتية	الاختصاص الدلالي للوحدات المعجمية مكانيًا/ زمانيًا/ ثقافيا
تحليل الأخطاء من حيث الأعمار/الخلفية الثقافية/ البيئة التعليمية إلخ.	الفيفي وأتويل al-Faifi and Atwell ٢٠١٤ (مدونة متعلمي العربية Arabic Learner Corpus)	معامل الارتباط بين نوع الأخطاء والخلفيات العمرية و/أو الثقافية و/أو العمرية لغير العرب

خاتمة

عُرِضَ آنفًا أهم الأدوات اللغوية التحليلية للنصوص العربية، ومفاهيم المدونة العربية الحاسوبية وأهم أنواعها، ومفهوم التصاحب اللفظي من وجهة نظر لسانية مدونيّة حاسوبية، وأهم الطرائق الإحصائية من حيث قياس القوى والمدى والتجاور والتضام، وأثر هذه القياسات الإحصائية في دعم الفرضيات اللغوية من عدمه. ولا غرو أن نلحظ أهمية هذه الاتجاهات في الدراسات اللغوية العالمية، واللغة العربية ليست بمنأى عن المستجدات البحثية اللغوية لخدمة لغة الضاد. ولقد وقف الباحث على ما يزيد عن ٤٠٠ دراسة أجنبية، وقد كانت أدوات وورد سميث WordSmith Tools منطلقَها الأساس في تحليل فرضياتها المتعددة. إن دخول المهتمين في اللغة والأدب إلى هذا المضمار سيبني أُطرًا جديدة لمزيد من التفسير والتحليل الآلي والإحصائي والنوعي الحذر، وسيضفي آفاقًا رحبة تعين على مزيد من استيعاب واقع اللغة العربية بوصفها لغةً حيّة مُنتجَة بالصور القياسية لمستويات العربية النموذجية كلها.

الحواشي

[٤٥] تتفرع المدونات إلى ما هو أكثر تفصيلًا من هذه التفريعات التي تتعلق بعدد لغة المدونة، وهنا أذكر -بشكل موجز- التقسيمات المعمول بها وفق النوع والغرض والعدد والتصميم إلخ. فعلى صعيد النوع، هناك مدونات لغوية حاسوبية للغات الكلاسيكية، وهناك ما هو للغة المعاصرة، وهناك ما يتضمن أوعية genres معينة مثل مدونة الصحف وهكذا دواليك. وثمة مدونات من حيث النصوص تتضمن نصوصًا ليست كاملة وتعرف بمدونات العينة sample corpus، ونصوصا حية كاملة من حيث الإنتاج، وتعرف بالمدونات كاملة النصوص full-text corpora. وهناك مدونات من حيث الوعاء الناقل للغة الحية، والذي

يكون نصا مكتوبا، أو منطوقا يُحول إلى نص مكتوب. وتفترق المدونات أيضا من حيث التنوع الزمني، فهناك مدونة لغوية معاصرة مثل مدونة السليطي al-Sulaiti (٢٠٠٤) (مدونة العربية المعاصرة Contemporary Arabic Corpus، انظر أيضا: السليطي al-Sulaiti 2009)، ومدونة زمنية متعلقة بالتطور الزمني اللغوي مثل مدونة هيلسنكي للنصوص الإنجليزية the Helsinki Corpus of English Texts التي تتضمن مليون ونصف المليون كلمة للإنجليزية القديمة والوسطى وبداية العصر الحديث (١٩٩١). كما تتمايز المدونات أيضا من حيث مستعملي اللغة إلى مدونات اللغة الأم ومدونات اللغة الثانية أو الأجنبية للمتعلمين (مثل مدونة أبوحكيمة Abuhakema 2009 ومدونة الفيفي وأتويل ٢٠١٢). أما من حيث الغرض، فهناك مدونات حاسوبية لأغراض عامة general-purpose corpora، ومدونات محددة المجال domain-specific corpora. أما التقسيم الثالث؛ فهو ما يتعلق بالعدد، أي: بعدد اللغة في المدونة، وتنقسم إلى مدونة أحادية وتقابلية ومتوازية. أما آخر صنف تتمايز به المدونات فهو ما يتعلق بالتصميم (بمعنى آخر: بالتوسيم النحوي Part of Speech Tagging والتحشية annotation) حيث إن هناك مدونات لم يعمل لها توسيم وتحشية، أو ناقصة من حيث عناصر التحشية: فواصل الصفحات، والفقرات، والفوارق والتشاكل الإملائيين، والأرقام، والزوائد، والرموز الأجنبية المختلفة عن رموز لغة المدونة الأصلية، وتواريخ النصوص، والمؤلفين، والأوعية، والمجالات، والموضوعات، والمستويات اللغوية، والسجل اللغوي (لمزيد من الأمثلة حول أنواع المدونات الغربية وأمثلتها؛ انظر: بيكر وآخرين Baker et al 2006: ٤٩، ٧١، ١٢٦، ١٤٢).

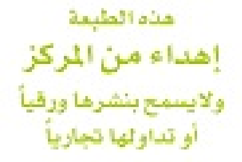
ووفقاً لهذا التصنيف، فإنه لا يمكن أن يقوم باحث لغوي -على سبيل المثال- بتحليل أنواع تراكيب الجمل الفعلية، أو وظائف حروف الجر الدلالية، أو تحليل الخطاب الديني في الثقافة العربية، أو تفسير الموجودات والمحسوسات والمواد الثقافية بعلم الوجود في اللغة العربية باستعمال مدونة عربية معاصرة (مثل

مدونة السليطي) أو باستعمال مدونة متعلمي اللغة العربية (مثل مدونة الفيفي وأتويل) إلا وأن يكون مدركاً بأية مدونة يكون لها أساس علمي يُجيب عن تلك الأنواع من التحليلات.

[٤٦] التوسيم النحوي Part-of-Speech Tagging (ويُسمى أيضا بـ Grammatical Tagging؛ انظر: ميكنري وآخرون McEnery et al 2006)، والتوسيم الدلالي Semantic Tagging، والتوسيم النبري Prosodic Tagging، والتحشية Annotation؛ كلها إجراءات حاسوبية مطلوبة لتطوير أي مدونة حاسوبية، وكل هذه التوسيمات، وبالأخص التوسيم النحوي، يصعب إنجازها في المدونة العربية الحاسوبية لسبب يُعزى إلى حاجتها إلى تطوير التوسيم النحوي بصورة يدوية لأقسام الكلم شكلًا ووظيفةً من أجل ضمان تطابق أكبر وأدق. ولو قلنا -مثلا-مدونة مكونة من مليون كلمة، فإن توسيم أقسام الكلام فيها يدويا يحتاج إلى فريق عمل يصل إلى المائة ربما، ويُنجز التوسيم بشكل يومي يدويا، وعلى مدى سنوات عديدة على أقل تقدير. وثمة توسيمات نحوية مقترحة وعديدة، ومعظمها يُمكن أن تحقق نسبًا مرضية من التطابق، ويرى الباحث أن نسب التطابق في المدونات العربية لا يُمكن أن تتجاوز نسبة ٧٠٪ للمدونات بسبب صعوبة ضبط كل وظائف الكلمات العربية المتنوعة بتنوع أنظمتها الخطية والصوائتية والتصريفية والنوعية والوظيفية. ويرى الباحث أيضا أن أي منجز أو اقتراح أو عمل تكون جهوياته منصبة على محاولات تطوير التوسيم النحوي في العربية، ولا يكون مقبولا بافتراضاته في اختباره وفحصه آليا، فإن نطاق تطوير التوسيم النحوي العربي سيكون منحسرًا ولن يُكتب له التطوير.

[٤٧] انظر إلى الحاشية رقم [٤٥].

[٤٨] ثمة مدونة متوازية ليست متوفرة، وصُممت من قبل جامعة جون هوبكنز John Hopkins University، وهي مدونة توازي مقوّمات constituents الآيات القرآنية بمقابلاتها الترجمية في الإنجليزية، أي أنها مدونة تتضمن آيات القرآن



الكريم وما يقابلها في الترجمة الفعلية؛ حيث لكل مقوم وعنصر لفظي في كل آية ما يوازيه (أو يحاذيه aligning؛ من محاذاة alignment) من مقابلات في اللغة الإنجليزية. وحيث إن هذه المدونة ليست متاحة، فلا يُعلم ما إن كانت النصوص المترجمة دقيقة للنص القرآني أم لا.

[٤٩] ينقسم معظم المدونات العربية الحاسوبية من حيث التوسيم إلى قسمين؛ الأول: شبه الموسّمة semi-tagged (بخلاف الموسّم جزئيًا partially tagged) كحال مدونة أرابيكوربوس ٠٠٠,٦٠٠,١٧٣ مليون كلمة arabicCorpus http:// arabicorpus.byu.edu، والمدونة العربية KACSTAC بنسختها الجديدة. أما القسم الثاني فهو غير الموسّم non-tagged كحال معظم المدونات العربية العامة، ومنها مدونات العربية الشبكية Arabic Internet Corpora (٣١٧ مليون كلمة) لـ سيرج شاروف Serge Shroff (انظر: /http://smlc09.leeds.ac.uk query-ar.html). ولا توجد مدونة عربية حاسوبية موسومة بصورة كاملة كحال، على سبيل المثال، مدونة بيرمنغهام الإنجليزية WordBanks Online (البنك الإنجليزي Bank of English http://wordbanks.harpercollins. co.uk/Docs/Help/guide.html). ثمة مدونات عربية تتضمن نصوصا عربية رقمية، ولا يُرى إضافتها هنا نظرًا لاختصاصها بسجلات لغوية وأوعية لغوية خاصة للعربية، هذا بالإضافة إلى عدم تجاوز عدد كلماتها ١٠٠ مليون كلمة (انظر: http://www.kacstac.org.sa/osact/UsefulResources.html)، ومنها المُجذّع جزئيا partially tokenized والموسم جزئيا partially tagged (وليس شبه الموسم semi-tagged): Arabic Gigaword Corpus Firth Edition (انظر الثبيتي al-Thubaity 2014).

[٥٠] انظر (بارلو Barlow 2014) /http://www.monoconc.com

[٥١] IT Servieces' Oxford University: برنامج يعالج النصوص وملحق بالمدونة الوطنية البريطانية British National Corpus؛ انظر: .http://www

natcorp.ox.ac.uk/tools/index.xml ويمكن تحميل البرنامج من هذا الرابط http://sourceforge.net/project/showfiles.php?group_id=١٣٠٢٨٩. ويدعم هذا البرنامج تحليل مصادر لغوية ضخمة ومدونات لغوية طبيعية على امتداد ملفات XML.

[٥٢] انظر (تسوكاموتو Tsukamoto n.d.) http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/index_e.html

[٥٣] انظر (روبرتس Roberts 2014) http://www.andy-roberts.net/coding/aconcorde

[٥٤] انظر (سكوت Scott 2012) WordSmith Version 6 http://www.lexically.net/wordsmith/version6/index.html

[٥٥] انظر الثبيتي وآخرين al-Thubaity et al (ACP Tool) 2013 http://sourceforge.net/projects/kacst-acptool

[٥٦] انظر: https://the.sketchengine.co.uk/login

[٥٧] هذه القياسات الإحصائية كافية فيما يراه باحث هذه الورقة، وهي أساس البحث اللغوي في أي مدونة لغوية حاسوبية. وثمة حزم أخرى لا تختلف عن هذه الأساس، ونفاط اختلافها تكمن في قراءات خوارزمية أخرى تُوصل إلى ما يقربُ من فوائد المعطيات الأساسية ذاتها (أي: قبول قوة التصاحب من عدمه).

[٥٨] يقوم هذا التقسيم أساسا على نقل مفاهيم ستيفان جريس Stefan Gries (٢٠١٠) حول إعادة نظرته إحصائياً لمصداقية التحليل الحاسوبي لأي مدونة، ويلخص من خلال تحليليه أن أية مدونة من حيث الأصناف الأربعة: النوع أو الغرض أو العدد أو التصميم، فإن على الباحث فيها أن يلتزم في تقديم

الإحصاءات بما يتوافق وعينة تلك الأصناف الأربعة، وبشكل دقيق فيما يخص التصاحب أو التلازم أو التجاور أو التضام (التقارب).

أما هذه المقابلات العربية فيُرى ضبطُها بصورة لعلها تكون مقبولة مستقبلا، وهي قابلة للدحض بلا شك؛ وشكل ضبطها هو على النحو الآتي: التصاحب collocation، والتلازم colligation، والتجاور collostruction، والتضام (الدلالي) semantic set. والمعاني اللغوية لكل مقابل عربي هي: مفهوم التصاحب (أي: القريب اللفظي المباشر) أو التلازم النحوي (القريب النحوي المباشر)، أو التجاور (التجاور اللفظي تركيبًا وغير المباشر)، أو التضام الدلالي (أي التقارب الدلالي semantic preference) أي: التضامات الدلالية التي تشترك في العلاقات الوجودية للحقول الدلالية ontological relationships of Arabic semantic fields (انظر الحاشية ٤٩). وهذا الأخيرة هي حلم الدرس اللغوي العربي، ويتحتّم أن تكون واقعًا مأمولًا لكل مختص عربي في الدرس اللساني الحديث، وعلى كل باحث لغوي أن يأخذ في تخصصه شيئا من التقنيات التي تُوفرها أدوات معالجة اللغة آليا، والكيفية النموذجية في بحث سيُكوّن مستقبلًا كمًا يقدم لقرائه مزيدًا من التفتّق المعرفي الذي يأتي به باحثون جدد حول هذا الموضوع. إن هذه القضية وما ذُكر هنا هو شرط العلوم البنائية في العلوم الحديثة.

[٥٩] الإشارة هنا إلى إحدى المدونات العربية المحوسبة التي تُمكّن نمطًا من أنماط البحث التمكينيّة؛ والمتوفرة في نظام الاستعلام query system، أو بالاستناد إلى مدونة عربية يقوم الباحث بجمعها في ملف text ومن ثمّ إضافتها إلى أداة من أدوات تحليل اللغة العربية وأفضلها غواص و Sketch Engine. وتُربط هذه الخصائص البحثية بثلاث مدونات عربية رئيسة، كل واحدة منها تُوفّر أدوات تحليلية خاصة بها. ولزاما على كل باحث لغوي أن ينظر إلى مستجدات تلك المدونات العربية الثلاث: المدونة العربية (مدينة الملك عبدالعزيز

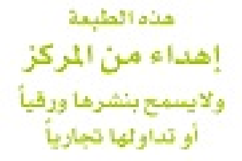
للعلوم والتقنية) http://www.kacstac.org.sa / ، ومدونة أرابيكوربوس (جامعة بريغهام يونغ الأمريكية) /http://arabicorpus.byu.edu، ومدونة جامعة ليدز لمحتوى العربي الشبكي من بعض الصحف وكامل الويكيبيديا .http://corpus. leeds.ac.uk/internet.html. وفي كل مدونة من هاته المدونات أدوات بحث خاصة تتطور، غير أن أكثر الأدوات الواعدة تطورًا فيها هي المدونة العربية، ومن ثمة مدونة أرابيكوربوس كونها قد زادت من عام ٢٠١٠ حتى عام ٢٠١٢ بما يزيد عن ٩٠ مليون كلمة مع إمكانية البحث عن الجذر اسميًا وفعليًا، وعن الجذع ولواصقه الاشتقاقية والتصريفية (word forms)، وعن إمكانية عرض النص الأصلي كاملًا بتطابقات الكلمة المبحوث عنها.

[٦٠] هذا الدراسة الأجنبية ودراسة الحمزاوي (١٩٨٦) من قبلها تُعدّان المحك الأساس والتحدي الأكبر لقضايا اللغة المتنوعة زمنيا ووصفيًا من جهة. ومن جهة ثانية: التحدي الآخر الذي يتطلّب من المؤسسات أن تجعل من التوقيف (أو التوسيط) للمتغيرات اللغوية استعمالا الإطارَ العام لـ: التوصيف الدقيق لمواد الثقافة العربية بالوحدات المعجمية العربية، والاستعمالات التركيبية التوليدية المستجدة، والوحدات المعجمية العربية المعربة والدخيلة، والأساسَ لها جميعًا في بناء مواد العربية التعليمية ومعاجمها متعددة الأغراض.

المراجع

المراجع العربية

الثبيتي، عبد المحسن وآخرون. غواص، KACST, ACP Tools، http:// sourceforge.net/error-404.html?project=kacst-acptool.



حبش، نزار. مقدمة في المعالجة الطبيعية للغة العربية، ترجمة: هند بنت سليمان الخلفية. دار جامعة الملك سعود للنشر، الرياض، ٢٠١٤.

الحمزاوي، محمد رشاد. العربية والحداثة أو الفصاحة فصاحات، دار الغرب الإسلامي، بيروت، ١٩٨٦.

العصيمي، صالح بن فهد. علم المتون وعلوم اللغة. مجلة كلية الآداب والعلوم الإنسانية، كلية الآداب والعلوم الإنسانية، فاس (المملكة المغربية)، العدد (١٩)، ٢٠١٣، ص. ٣٧-٦٧.

الفيفي، عبدالله وأتويل، إريك. المدونات اللغوية لمتعلمي اللغة العربية: نظام لتصنيف وترميز الأخطاء اللغوية. المؤتمر الدولي لعلوم وهندسة الحاسوب باللغة العربية (الدورة الثامنة)، ٢٦-٢٨ ديسمبر، جامعة القاهرة، ٢٠١٢.

المراجع الإنجليزية

Abuhakema, G., Feldman, A. and Fitzpatrick, E. ARIDA: An Arabic Interlanguage Database and Its Applications: A Pilot Study. Journal of the National Council of Less Commonly Taught Languages, 7, 2009, pp. 161-184.

Al-Ajmi, H. Compiling an English-Arabic parallel text corpus. In: Proceedings of Asian Association for Lexicography, August 27-29, Urayasu (Japan): Meikai University, 2003, pp.51-54.

Al-Faifi, Abdullah and Atwell, Eric. Arabic Learner Corpus and Its Potential Role in Teaching Arabic to Non-Native Speakers. In the Proceedings of the Seventh Biennial IVACS conference, 19 - 21 Jun 2014. Newcastle (UK), 2014.

Al-Mujaiwel, S. Contrastive Lexicology and Comparable English-Arabic Corpora-Based Analysis of Vague and Mistranslated Arabic Equivalence. Ph.D. thesis, Exeter University, 2012.

Al-Sulaiti, L. Designing and Developing a Corpus of Contemporary Arabic, MSc dissertation. Leeds: Leeds University's School of Computing, 2004.

Al-Sulaiti, L., A Survey of Arabic Corpora. 2009. http://www.comp.leeds.ac.uk/eric/latifa/arabic_corpora.htm

Al-Thubaity, A., et al. New Language Resources for Arabic: Corpus Containing More Than Two Million Words and a Corpus Processing Tool. In the Proceedings of the International Conference on Asian Language Processing in Urumqi, China, 2013.

Al-Thubaity, et al. ACP Tool. Available for free use in: http://sourceforge.net/projects/kacst-acptool/, 2013.

Al-Thubaity, A. O. A 700M+ corpus: KACST Arabic corpus design and construction. Language resources and evaluation, 2014. DOI 10.1007/s10579-014-9284-1. Arabic

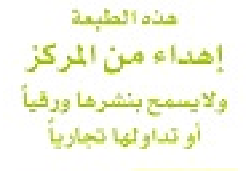
Baker, P. et al. A Glossary of Corpus Linguistics. Edinburgh: Edinburgh University Press, 2006.

Baker, P. Sociolinguistics and Corpus Linguistics. Edinburgh University Press, 2010.

Barlow, M. MonoConc, version 2.2, 2014. http://www.monoconc.com/.

Bernard, Lou and Dodd, Tony. Xaira: an XML aware tool for corpus searching. In: Archer, D., Rayson, P., Wilson, A., and McEnery, T., eds., the Proceedings of the Corpus Linguistics Conference, 16, 2003, pp. 142–144, UCREL, University of Lancaster.

Berry-Rogghe, Godelieve. The Computation of Collocations and Their Relevance in Lexical Studies. In: Aitken, A., J., Bailey, R., W., and Hamilton-Smith, N., the Computer and Literary Studies. Edinburgh University Press, 1973, pp. 103-112.



Church, Kenneth, and Hanks, Patrick. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics 16(1), 1990, pp., 22-29.

Diab, Mona. Improved Arabic Base Phrase Chunking with a New Enriched POS tag set. In Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, 2007, pp. 89-96.

Diab, Mona. Second Generation AMIRA Tools for Arabic Processing: Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking. In: Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, 2009.

Firth, J., R. A Synopsis of Linguistic Theory 1950-1955: Studies in Linguistic Analysis. Blackwell, Oxford, 1957.

Goldberg, Adele E. The Nature of Generalization in Language. Cognitive Linguistics, 20(1), 2009, pp. 93-127.

Granger, S. et al., eds. Corpus-based Approaches to Contrastive Linguistics and Translation Studies, Rodopi, Amsterdam, 2008.

Granger, S. The Corpus approach: a common way forward for Contrastive Linguistics and Translation Studies. In: Granger, S. Lerot, J. and Petch-Tyson, S., eds., Corpus-based Approaches to Contrastive Linguistics and Translation Studies. Rodopi, Amsterdam, 2008, pp. 18-29.

Gries, St. Th. Collostructions: investigating the interaction between words and constructions. International Journal of Corpus Linguistics, 8(2), 2003, pp. 209-243.

Gries, St., Th. “Useful Statistics for Corpus Linguistics.” In: A. Sánchez and M. Almela, eds., a Mosaic of Corpus Linguistics: Selected Approaches. Frankfurt: Peter Lang, 2010, pp. 269–291.

Gries, St., Th. Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics, 13(4), 2008, pp. 403–437.

Gries, St., Th. Quantitative Corpus Linguistics with R: A Practical Introduction, Routledge, London, 2009.

Habash, Nizar, et al. MADA+TOKEN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In Proceedings of the second International Conference on Arabic Language Resources and Tools, Cairo, 2009.

Hunston, S. Corpora in Applied Linguistics. Cambridge: Cambridge University Press, 2002.

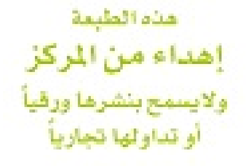
Hunston, Susan., Colligation, Lexis, Pattern, and Text. In: Scott, Mike, and Thompson, eds., Patterns of Text: In Honour of Michael Hoey. John Benjamins, Amsterdam, 2001, pp. 14-33.

IT Services's Oxford University. British National Corpus and Xaira, http://projects.oucs.ox.ac.uk/xaira/ and http://www.natcorp.ox.ac.uk/tools/index.xml.

Jakubíček, M., et al. Fast syntactice searching in very large corpora for many languages. PACLIC 2010, Japan. http://www.sketchengine.co.uk/documentation/wiki/SkE/DocsIndex.

Jakubíček, Miloš, et al. Fast syntactic searching in very large corpora for many languages, Japan, PACLIC, 2010, pp. 741-746.

Johansson, S. Contrastive linguistics and corpora. In: Granger, S. Lerot, J. and Petch-Tyson, S., eds., Corpus-based Approaches to Contrastive Linguistics and Translation Studies. Rodopi, Amsterdam, 2008, pp. 31-44.



Kilgariff, Adam, et al. A quantitative Evaluation of Word Sketches. EURALEX, the Netherlands, Leeuwarden, July 2010.

Kilgarrif, Adam, et al. The Sketch Engine (Lexical Computing Ltd.), https://the.sketchengine.co.uk/login/.

Kilgarrif, Adam, et al. The Sketch Engine: ten years on. Lexicography, 1(1), 2014, pp. 7-36.

Kilgarriff, Adam, et al. The Sketch Engine. In: Proceedings of EURALEX, Lorient, France, 2004, pp. 105-116, http://www.sketchengine.co.uk.

Lipka, L. English Lexicology: Lexical Structure, Word Semantics, and Word-Formation. Tübingen: Narr Studienbücher, 2002.

McEnery, Anthony. Corpus Linguistics. In: Mitkov, R., ed., Oxford Handbook of Computational Linguistics. Oxford: Oxford University Press, 2003, pp. 448-463.

McEnery, Tony and Hardie, Andrew. Corpus Linguistics. Cambridge University Press, Cambridge, 2012.

McEnery, Tony and Xiao, Zhonghua. Parallel and comparable corpora: What is happening? In: Anderman, G. and Rogers, M., eds., Incorporating Corpora: The Linguist and the Translator. Clevedon: Multilingual Matters, 2007, pp. 18-3.

McEnery, Tony, et al. Corpus-Based Language Studies, an advance resource book. Routledge, Oxford, 2006.

Moll, Clàudia, P. When diachrony meets synchrony. How linguistic variation sheds light on the origin of synchronic phonological processes and theories of language change. Dialect Laboratory. Dialects as a testing ground for theories of language

change. Studies in Language Companion Series, 128, 2012, pp.197-225.

Oakes, M. Statistics for Corpus Linguistics, Edinburgh: Edinburgh University Press, 1998.

Piao, S., Sentence and word alignment between Chinese and English. Ph.D. thesis, Lancaster University, 2000.

Piao, S. Word alignment in English-Chinese parallel corpora. Literary and Linguistic Computing 17(2), 2002, pp. 207-30.

Price, Todd L. Structural Lexicology and the Greek New Testament: Applying Corpus Linguistics for Word Sense Possibility Delimitation Using Collocational Indicators. Ph.D. thesis. Middlesex University, 2013.

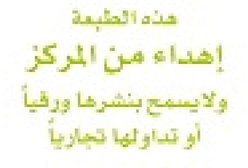
Roberts, A., al-Sulaiti, L., Atwell, E. aConCorde: Towards an open-source, extendable concordancer for Arabic. Corpora, Edinburgh University Press, 1, 2006, pp. 39-57.

Roberts, Andrew. aConCorde. 2014. http://www.andy-roberts.net/coding/aconcorde.

Rouhani, Jameela. An Applied Research Into the Linguistic Theory of Collocation: English-Arabic Dictionary of Selected Collocations and Figurative Expressions with an Arabic Index. Ph.D. thesis, Exeter University, 1994.

Rychly, Pavel. A Lexicographer-Friendly Association Score. In: Proceedings of 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN, Masaryk University, Brno, 2008.

Scott, Mike. WordSmith Tools 5.0. Lexical Analysis Software, 2010.



Scott, Mike. WordSmith Tools, version 6. Lexical Analysis Software, Liverpool, 2012. http://www.lexically.net/wordsmith/version6/index.html.

Sharoff, S. Harness¬ing the lawless: using comparable corpora to find translation equiva¬lents. Journal of Applied Linguistics, 2004, 1(3), pp. 333-350.

Sinclair, J. Corpus, Concordance, Collocation. Oxford University Press, 1991.

Sinclair, J. Reading Concordances: An Introduction. London: Pearson, 2003.

Sinclair, J. Trust the Text: Language, Corpus and Discourse. Edited with Ronald Carter. Routledge, London, 2004.

Tognini-Bonelli, E., Lexis in contrast. In: Granger, S. and Altenberg, B., eds., Studies in Corpus Linguistics. Benjamins, Amsterdam, 2000, pp. 3-48.

Tsukamoto, Satoru. KWIC Concordance, (n.d.) http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/index_e.html.

خاتمة

نظرة تأمل الماضي واستشراف المستقبل

صالح بن فهد العصيمي

خاتمة: نظرة تأمل الماضي واستشراف المستقبل:
صالح بن فهد العصيمي

إن المدونات اللغوية ولسانيات المدونات اللغوية حقل جديد نسبياً في مجال العمل اللساني والدرس اللغوي (محمود: المبحث الأول). وقد عرضنا في مباحث هذا الكتاب بعض الميادين التي طرقها هذا العلم في مجال البحث اللغوي الحديث. وعلى الرغم من تعدد هذه الأبحاث والموضوعات والميادين فلا زال المجال واسعاً للاستفادة من المدونات في مجالات أخرى وباستخدام طرائق شتى. ومن المهم الإشارة إلى أن العمل في مجال المدونات العامة أو المتخصصة كمدونات المتعلمين العربية، ووسمها، والاستفادة منها في مجالات البحث اللغوي - أو غيرها من المجالات - لا يزال في بداياته، ويحتاج للكثير من الدراسات ليصل إلى مستويات أفضل (الفيفي: المبحث الثاني)، وتليق بلغة عالمية كالعربية. إن دراسة العربية المعاصرة، وقضايا الأساليب المستجدة فيها، والدراسات المتعلقة بتحليل الخطاب المعاصر لهي من الموضوعات التي ينبغي أن يلتفت إليها اللغويون العرب مسلحين بالأدوات المنهجية للسانيات المدونات (الشمري والثبيتي: المبحث الرابع).كما أن دراسة المفاهيم المعجمية مثل التصاحب اللغوي وغيره سيعيد رسم الأبعاد الاستعمالية للوحدات المعجمية في المعجم الذهني العربي (المجيول: المبحث الخامس). وكذلك فإن دخول المهتمين باللغة والأدب وفروع اللسانيات العربية التقليدية إلى مضمار لسانيات المدونات سيبني أُطرًا جديدة لمزيد من التفسير والتحليل الآلي والإحصائي والنوعي الحذر، وسيضفي آفاقًا رحبة تعين على مزيد من استيعاب واقع اللغة العربية بوصفها لغةً حيّة مُنتجَة بالصور القياسية لمستويات العربية النموذجية كلها (المجيول: المبحث الخامس).

إن هناك حاجة ملحّة لإنشاء مدونات متعددة للغة العربية مثل مدونة التلاميذ (مكتوبة ومنطوقة) في مراحل الاكتساب اللغوي والتعليم العام حتى يتسنى

للغويين دراسة الضعف اللغوي وتقديم الحلول حوله. ومثل ذلك في الأهمية إنشاء المدونات المتخصصة مثل الأكاديمية المنطوقة والمكتوبة، ومدونات الخطاب القضائي والسياسي والدعوي وغيرها من المجالات المتخصصة. كما أن هناك ضرورة قصوى لأن يعتمد المعجم العربي على المدونات، وأن يتعدد المنتَج المعجمي بتعدد مجالات المعاجم مثل معاجم التصاحب والترادف والتضاد وغيرها. بالإضافة إلى ذلك فدراسة العاميات واللهجات العربية دراسة قائمة على منهج لسانيات المدونات مهمة إذا أردنا المقارنة بين الفصحى ومستوياتها حتى نتمكن من توصيف أعمق لهذه اللغة الممتدة على ما يزيد عن أربعة عشر قرناً.

إن مما ينبغي التأكيد عليه أن مجالات لسانيات المدونات وإن كانت تعتمد على التقنية التي اختصرت الجهد والوقت والمال فإنه لا غنى عن الفِرَق البحثية والعمل المؤسسي نظراً لضخامة العمل الذي ينتظر اللغة العربية في مجالاتها اللسانية الحديثة.

ومما يبعث على التفاؤل أن هناك انطلاقة حقيقية يشهدها الدرس اللغوي العربي في مجال لسانيات المدونات، ويوجد حالياً مختصون في هذا المجال منهم الباحثون المشاركون في هذا الكتاب أثروا هذا الحقل الحديث بجهودهم الفردية، ومنهم الباحثون طلاب وطالبات الدراسات العليا في علم اللغة التطبيقي في جامعة الإمام محمد بن سعود الإسلامية الذين بدؤوا يتجهون إلى هذا النوع من الدراسات، بالإضافة إلى باحثين آخرين على امتداد العالم العربي وكذلك في العالم الغربي والمراكز البحثية التي تعنى بهذا النوع من الدراسات. وقد جاء هذا الكتاب الثري ليقدّم استعراضاً شاملاً للسانيات المدونات إلا أن هناك فرصة أخرى لإصدار كتاب آخر يعالج القضايا المتقدمة والمسائل الأعمق والتطبيقات الأوسع في مجال لسانيات المدونات، وهو ما نؤمله من مركز الملك عبدالله الدولي لخدمة اللغة العربية ومن غيره من الجهات العلمية والأكاديمية في العالم العربي عامة وفي المملكة العربية السعودية خاصة.

مسرد المصطلحات

تحشية إحالية	anaphoric annotation
توسيم إحالي	anaphoric tagging
تحشيةبإضافة معلومات داخل النص	annotation
محاذاة آلية	automatic alignment
قيمة محسوبة	calculated value
مربع كاي	chi square
تلازم نحوي	colligation
تصاحب	collocation
نظرية التصاحب	collocation theory
مؤشر التصاحب	collocational indicator
تلازم معجمي-نحوي	collostruction
مدونات متقابلة (مقارَنة)	comparable corpora
مكملات (متممات)	complements
لغة وسيطة	compromise language
اللسانيات الحاسوبية	computational linguistics
صرفيمات سلسلية	concatenative morphemes
سطور الكشاف (المِكشاف) السياقي	concordance line
أداة ربط	conjunction
قواعد التركيب	construction grammar
مدونة (متن)	corpus
لسانيات المدونات (المتون)	corpus linguistics

corpus query language	لغة استعلام المدونة
corpus-based approach	منهج (مذهب) معتمد على مدونة
corpus-based research	بحث معتمد على مدونة
corpus-driven approach	منهج (مذهب) موجَّه بمدونة
corpus-driven research	بحث موجَّه بمدونة
Corpus Encoding Standard: CES	المرجع القياسي لترميز المدونات (معيار تشفير المدونة)
critical value	قيمة واقعية
degree of frequency	درجة التكرار
derivational suffix	لاصقة اشتقاقية
diachronic	تعاقبي
diachronic variation	تنوع تعاقبي
domain	مجال (نطاق)
electronic text analysis	تحليل نصوص إلكترونية
expected frequencies	تكرارات متوقعة
frequency	تكرار
General Standard Markup Language: GSML	لغة التعليم المعيارية العامة
genre	نوع
grammatical tagging	توسيم نحوي
Historiography	تاريخ تحليلي
homonym	المشتركات اللفظية
inflectional suffix	لاصقة تصريفية

معامل كسب المعلومات	Information Gain
تداخل اختصاصي (تداخل تخصصات متعددة)	interdisciplinary
تماثل معجمي	isolexical
تماثل نصي	isotexutal
هندسة اللغة	language engineering
مصادر اللغة	language resources
مدونة ضخمة	large-scale corpus
تجريد الكلمة من الزوائد	lemmatization
مواد الكلمة المعجمية	lexical word-items
لوج دايس	logDice
معامل الاحتمالية اللوغاريثمي	Log Likelihood
تعليم (من العلامة)	Mark-up
معامل المعلومات المتبادلة	mutual information
لغة طبيعية	natural language
معالجة اللغة الطبيعية	natural language processing
المادة العنقودية	nodal item
صرفيمات غير سلسلية	non-concatenative morphemes
فرضية صفرية	null hypothesis
تكرارات ملحوظة (مُلاحظَة)	observed frequencies
علم الوجود	ontology
مدونات سانحة	opportunistic corpora
مدونة متوازية	parallel corpus

parsing	تحليل نحوي
part-of-speech tagging	توسيم نحوي
polysem	البوليزيمي (كلمة ذات معانٍ متعددة)
prosodic tagging	توسيم نبري
query engine	محرك استعلام
query system	نظام استعلام
register	سجلّ
search engine	محرك بحث
semantic preference	تقارب دلالي
semantic set	مجموعة دلالية
semantic tagging	توسيم دلالي
span	مدى
standard level	مستوى نموذجي (معياري)
statistical corpus linguistics with R	لسانيات المدونات الإحصائية بـ «آر»
structural lexicology	علم المعاجم التركيبي
synchronic	آني (تزامني)
synchronic variation	تنوع آني (تزامني)
syntagmatic lexical units	وحدات معجمية نظمية
Text Encoding Initiative(TEI)	مبادرة ترميز النصوص
thematic structure	بنية صورية (موضوعاتية)
token	الكلمة الفعلية

tokenization	تجذيع
treebanks	البنوك الشجرية
t-score	قياس-ت
type	الكلمة النوعية
variation	تنوع
wildcard character	المحرف البديل
z-score	قياس-ز

فهرس